SCIENTIFIC AMERICAN

"IMPRINTING" IN NATURE

ONE DOLLAR

August 1972

1972 Pinto Wagon, shown at top with Squire option, optional whitewall tires, and luggage rack; shown at bottom, with optional whitewall tires and Base Accent Group.

The Pinto Wagon. For people who'd like an economy car if it carried more. Or a wagon if it cost less.

It's a solid Ford Pinto.

Under the hood, you'll find a tough little 2000cc overhead cam engine as standard equipment.

The body is welded solid. Then electrocoated against corrosion. Then painted five more times.

There's rack and pinion steering like on some expensive sports cars.

In normal use, recommended maintenance is only every six thousand miles or every six months.

So if it's economy you're after, you'll find plenty in this new Pinto.

It's a convenient Ford Wagon.

You can get it as a basic wagon (below), or with the Squire option (above).

It's only 10 inches or so longer than our Pinto Sedan: easy to handle and park. But put the rear seat down and you get over 60 cubic feet of cargo space. VW Squareback and Vega Kammback both give you about 50 cubic feet.

The lift gate swings up out of the way. The spare tire has its own well under the floor. The rear seat passenger windows flip open for ventilation.

In short, the Pinto Wagon gives you the kind of convenience and usefulness you'd expect from any Ford Wagon. Only it gives it to you in a basic little Pinto size.

When you get back to basics, you get back to Ford.

FORD PINTO

FORD DIVISION Ford

IF WE GOT A CALL TODAY AT 9AM TO BUILD A FLOATING CITY, WE COULD GO TO WORK ON IT AT 9:05.

Nowhere is the problem of population growth more evident than in coastal cities.

Tokyo. New York. San Francisco. Los Angeles. And Hong Kong. And these cities will run out of land. But there is a solution.

Floating cities.

Totally contained environments to relieve urban congestion.

And we could go to work to build them today.

Because we've literally built floating cities for more than a generation. Aircraft carriers. We've built 20.

And the two nuclear-powered carriers now under construction will each have a crew of about 6,000.

So you see, we have the expertise.

Because we're already at work on marine systems, enclosed environments, energy production and industrial construction.

Perhaps the first floating city would only have a population of 6,000. But ultimately, floating cities could be equal to Chicagos on the seas.

Exciting? Yes. Out of sight? No.

Floating cities are just around the corner. And when the time comes to build them, we're ready.

Newport News Shipbuilding (TENNECO) **Where tomorrow is. Now.**

A Tenneco Company Newport News, Virginia 23607

SCIENCE/SCOPE

A "four-for-four" record for the Intelsat IV was achieved recently when the fourth of the giant communications satellites was successfully put in orbit over the Indian Ocean. It joins the Intelsat IVs already stationed over the Atlantic and Pacific to complete the global network. The newest satellite increases by five times the communications capacity among 17 nations in Africa, Asia, Australia, and Europe. It was built by Hughes and an international team of subcontractors for Comsat, manager for the 83-nation Intelsat Consortium.

A small hand-held surgical instrument that uses space-age thermal transfer techniques to bring cryogenic temperatures directly to a surgical area was introduced at the 71st Annual American Proctologic Society Conference in New York recently. Called the Kryostik™, it consists of a 12-inch probe and a ½-liter reservoir that can sustain the probe tip at temperatures lower than -190°C for over 30 minutes. It will be manufactured by Hughes and marketed worldwide by the Ritter Company, division of Sybron Corporation.

The U.S. Air Force's F-15 air superiority fighter, rolled out in St. Louis recently by McDonnell Douglas Corp., will have an attack radar system developed by Hughes which acts as an electronic extension of the pilot's eyes and mind. It automatically scans and acquires targets, makes complex computations, and displays on the cockpit windscreen the instant information he needs for successful air-to-air combat. The new radar makes the F-15 the only USAF fighter that can locate and track low-flying aircraft in the ground "clutter" that blinds conventional radars.

A new high-temperature strain gage for testing aircraft, missiles, and space vehicles in a simulated flight environment has been developed for the U.S. Air Force by Hughes. It can measure structural stresses imposed at temperatures up to 2000°F, producing capacitance instead of resistance changes as a measure of strain. Only ½-inch long, it can be welded or bonded to airframe surfaces. DoD and NASA have bought several hundred of the gages for use in such programs as the space shuttle.

Hughes has immediate openings for engineers in the following categories: field engineers, technical instructors, ground support equipment engineers, test equipment design engineers, technical writers, and configuration engineers. Must have at least a BSEE degree. All openings require U.S. citizenship. Please write: Mr. H. G. Staggs, Hughes Aircraft Company, Field Service & Support Division, P.O. Box 90515, Los Angeles, CA 90009. Hughes is an equal opportunity M/F employer.

Microwave filters for communications satellites and other spacecraft must be made of material with very low thermal expansion. So far, the only successful material has been invar, a very heavy alloy of iron and nickel. On some recent satellites, as much as 120 pounds of invar filters have been used. Now Hughes has developed prototype microwave filters made of graphite epoxy composite materials. They have performed better than their invar counterparts, weigh only 20 to 40 percent as much, are much easier to manufacture, and are expected to cost less.

Established 1845 **SCIENTIFIC AMERICAN** August 1972 Volume 227 Number 2

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $10 PER YEAR. EUROPEAN SUBSCRIPTION COPIES DISTRIBUTED C/O INTERNATIONAL DISTRIBUTORS, VLAGSTRAAT 128, BORGERHOUT 2200, BELGIUM.

THE COVER

The painting on the cover shows a microphone next to eggs in the nest of a wild mallard. Sounds made by ducklings inside the eggs are picked up by the sensitive microphone and relayed to a tape recorder. Observation and recording of undisturbed natural events in the nest are carried out to provide a control situation for later experimental manipulation of "imprinting": the process by which newly hatched birds form an immediate and permanent attachment for the parent. Some major differences between the results of previous imprinting studies in artificial laboratory conditions and imprinting as it occurs in nature are discussed in this issue by Eckhard H. Hess, who has combined laboratory and field techniques to investigate the relative contributions of both prehatching and posthatching experiences on imprinting in mallards (" 'Imprinting' in a Natural Laboratory," page 24).

THE ILLUSTRATIONS

Cover painting by Lorelle M. Raboni

Page	Source
15	Phillip Trager
16–23	Dan Todd
24	Tom Prentiss
25	Eckhard H. Hess, University of Chicago
26	Tom Prentiss
27	Ilil Arbel
28	Tom Prentiss
29	Ilil Arbel
30	Tom Prentiss
31	Eckhard H. Hess, University of Chicago
33–40	Enid Kotschnig
48–50	Hale Observatories
51	Bart J. Bok, University of Arizona
52–54	Ilil Arbel
55	Bart J. Bok, University of Arizona
56	Hale Observatories
57–61	Bart J. Bok, University of Arizona
67	Paul Weller
69–74	Allen Beechel
76	F. G. Heath, Heriot-Watt University
77	Ben Rose
78	Lee Boltin (*top*); Fairchild Semiconductor (*bottom*)
80	F. G. Heath, Heriot-Watt University
81	Ilil Arbel
82	Paul Weller from *La Lumière Électrique*, Volume 6, 1882
85–95	Bunji Tagawa
97–102	Jim Egleson
103–104	Ira Pastan, Laboratory of Molecular Biology, National Institutes of Health
106–109	Ilil Arbel
110–113	Roger Hayward

THE BIG CAR INSIDE OUR LITTLE CAR.

Into each Vega a little Impala goes.
Power ventilation system.
Front disc brakes.
Side-guard door beams.
Two steel roofs instead of just one.

And, strange as it may seem, Vega's room-per-passenger compares pretty well with Impala's.

But Vega is no big car.
It's made for four people.

And although Vega's low center of gravity and wide stance make it feel bigger than a Vega, it's all done on a modest eight-foot wheelbase. (Two feet less than Impala's.)

Although our overhead-cam aluminum block engine is bigger than most little car engines, it's still awfully stingy with a gallon of gas.

And although Vega, with full foam seats and Full Coil suspension, is surprisingly comfortable to ride in from city to city, it's also a nifty car to zip around in from store to store.

Vega is a little car.
The little car that does everything well.

VEGA
CHEVROLET

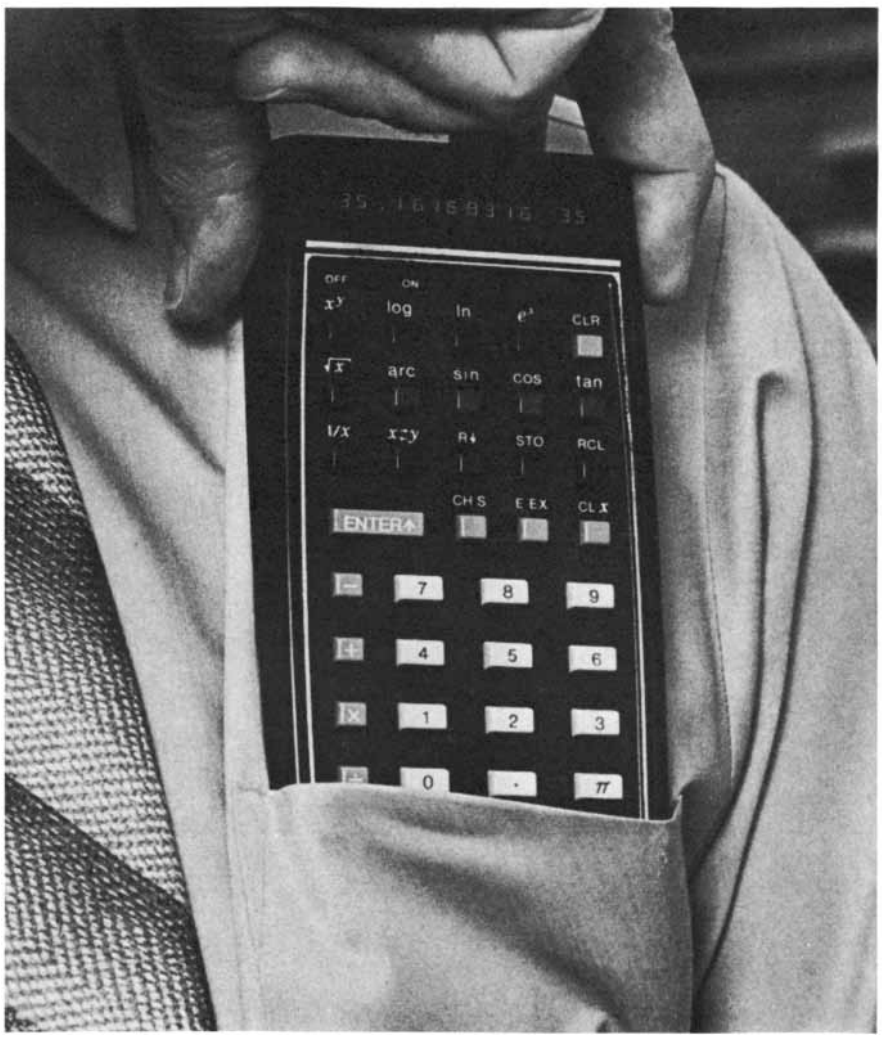
GM
MARK OF EXCELLENCE

Chevrolet. Building a better way to see the U.S.A.

Highway safety begins at home. Buckle up before you leave.

Some things are changing for the better.

Many people know us as an instrument manufacturer: we make more than 2,000 products for measurement, test and analysis. Others know us as a computer company: more than 10,000 own our programmable calculators and computers. We prefer to think that our business is to serve measurement, analysis and computation needs . . . in science, industry, medicine and education. That is the rationale behind every new instrument, computer or system that we tell you about in these ads. This month:



The HP-35 Shirt Pocket Calculator lets you make complex calculations like this one approximately five times faster than with your slide rule . . . with 10 place accuracy . . . and without a single scratch note!

$$\alpha_H = 1 + [3\times10^{-2}](10^{7.21}) + [8.7\times10^{-3}](10^{2.16})$$

$$pH = -LOG\sqrt{\frac{1}{\alpha_H}\left(\frac{3\times10^{-2}}{10^{11.7}} + \frac{8.7\times10^{-3}}{10^{7.21}}\right)}$$

**Chemists will recognize this as a calculation of the pH of a buffer solution for the mixture of Na_2HPO_4 @0.03 M/ℓ. and NaH_2PO_4 @8.7 x 10^{-3} M/ℓ.*

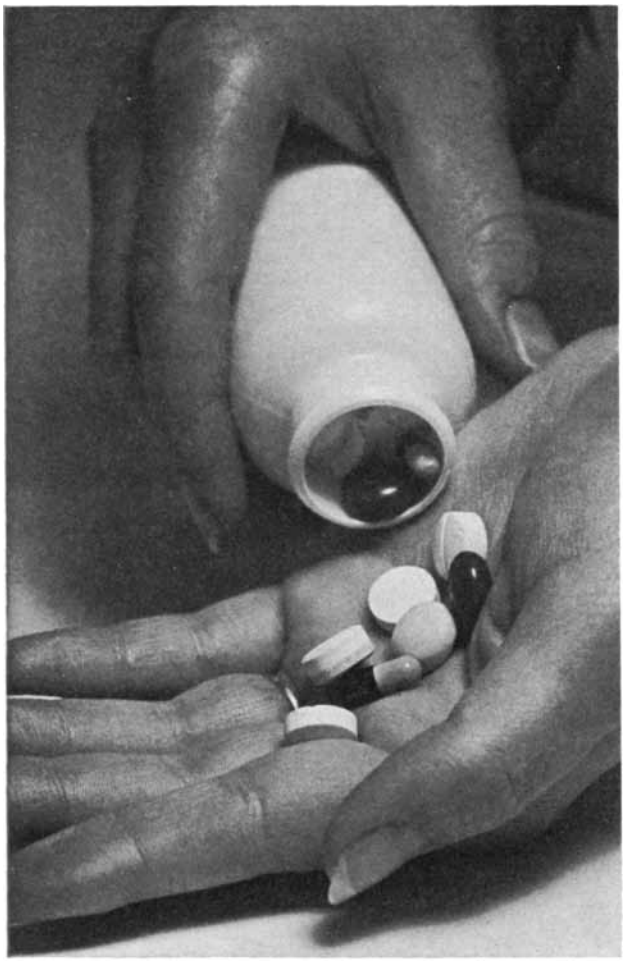
The new HP-35 Pocket Computer: a boon for scientists, engineers, or almost anyone.

When you first hear about it, it sounds like an electronic slide rule, the kind that's been available only in science fiction. Although it's only 3 by 6 inches and weighs 9 ounces, with rechargeable battery, it computes transcendental functions with a single keystroke, in less than a second. It calculates positive and negative numbers in floating point or scientific notation, automatically keeps track of the decimal throughout its 200-decade range, and displays answers that are accurate to the 10th significant digit.

And when you use it, you soon realize that it is more like a computer than a super slide rule. The secret is its 4-register operational stack. Solidly based on computer theory, it holds intermediate answers in a higher register and *automatically* brings them back when they are required for further calculations. The calculator also has a fifth register that lets you store any number and recall it to the working register at the touch of a key.

The power of our little wonder is illustrated in the calculation shown above. The HP-35 solves this problem in about 60 seconds and displays the answer to 10 significant digits, without a single scratch note.

The new HP-35 Pocket Calculator contains the equivalent of 30,000 transistors in specially designed MOS/LSI circuits. Yet it costs just $395 (domestic US price only). You may have to wait a while because demand has been so great. But if your people need this kind of computation power, it's worth the wait. Just use the coupon and we'll send full information to you.

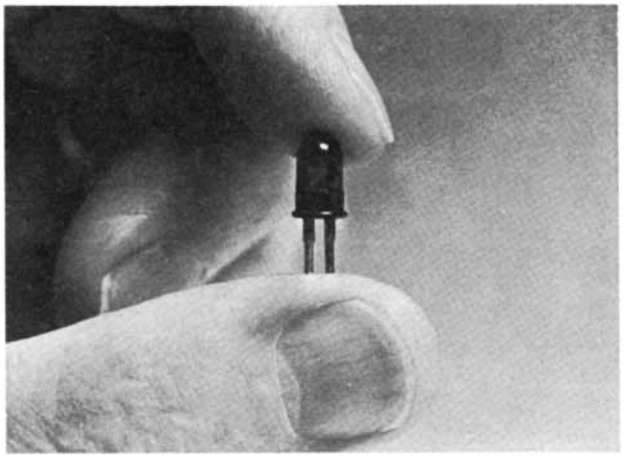


A faster, more efficient way to analyze drugs.

Not much can happen – neither emergency treatment for a drug overdose victim, nor prosecution of the pusher – until the drug has been positively identified. In large cities, where drug emergencies often reach into the hundreds daily, the chemist faces an enormously difficult problem, especially with traditional methods of chemical analysis. But there is a better way.

A laboratory in Charlotte (N.C.) recently sent us some powder from a confiscated pill for analysis on the new HP Gas Chromatograph/Mass Spectrometer/Computer System. Twenty minutes later, the analysis was complete: the pill contained heroin, morphine, and barbituric acid.

Fast, complete and positive, the analysis satisfied all medical and legal requirements. The HP system also takes a load off the lab's scientific staff because it can be successfully operated by technicians who have no special knowledge of mass spectroscopy or computers. The computer itself controls the operation of the spectrometer and records the mass spectrum while it makes all the necessary calculations, automatically. It can also compare the results of the analysis against a taped library of suspected components (in this case, a library of the mass spectra of 100 dangerous drugs) and automatically identify each of the sample constituents by name, positively. Where drugs are involved, that's an essential requirement. Just check the coupon for full information.



Try our solid-state lamp, free.

When Thomas Edison invented it, the incandescent lamp was one of the brightest ideas ever. It still is, if your object is illumination. But for display or indication, the new solid-state lamps (LED's) win hands down: they last at least 10 years and use only a few milliamps of current.

Because we are convinced that hands-on experience with an LED surpasses any description, we'd like to send you one of our 5082-4400 Series free. Use it even once, and we'll wager you'll never use any other kind of display lamp again.

Our LED is a gallium arsenide phosphide diode that's hard to ignore. You can see its red glow several yards away at any angle to 180°. It lights up when you connect a mere 1.6 volts to its rugged leads, draws only 2 to 10 milliamps, and includes a clip for easy panel-mounting.

Use the coupon to send for your free LED.* If it turns you on, we're ready to ship quantity orders immediately.

For more information on the calculator, our systems or a free LED fill out the coupon and send to: Hewlett-Packard, 1508 Page Mill Road, Palo Alto, California 94304; Europe: P.O. Box 85, CH-1217 Meyrin 2, Geneva, Switzerland; Japan: YHP, 1-59-1, Yoyogi, Shibuya-Ku, Tokyo, 151.

*OFFER EXPIRES OCTOBER 31, 1972 00213

LETTERS

Sirs:

Martin Davis and Reuben Hersh, in their article "Nonstandard Analysis" in the June issue of *Scientific American*, remark: "The atomism of Democritus had been meant to refer not only to matter but also to time and space. But then the arguments of Zeno had made untenable the notion of time as a row of successive instants, or the line as a row of successive 'indivisibles.'" The historical relationship between the ideas of Zeno and Democritus, however, seems to have been just the reverse. Zeno's paradoxes were directed not against infinitesimals as such but rather against infinite divisibility, that is, against the continuity of what we call today the "real line." Democritus, who most likely formulated his theory in *response* to Zeno, posited atoms that were not strictly infinitesimals but on the contrary were extremely small but finite physical entities whose aggregate composed the objects of the natural world. Democritus thus accepted Zeno's arguments and sought to evade the paradoxes of infinite divisibility by rejecting the possibility that division could in principle be carried out without limit. The problem was that Democritus at the same time desired that his atoms be partless, a property they were to have in common with geometrical points. Epicurus, possibly following Democritus but more likely taking his cue from Aristotle's *Physics*, defended a fully granular or quantized view of space and time, but he too seems to have posited elements that are finitely small in that their (finite) sums yield finite quantities, but that in other respects behave as points, presumably so that logic and geometry can retain their validity for the physical world. The details of these theories are still under investigation, but in general the problem the atomists had seems similar to the one that Davis and Hersh have described for more recent times, when the distinction between mathematics and physics was not strictly observed.

DAVID KONSTAN

Wesleyan University
Middletown, Conn.

Sirs:

The review of the book on Christo, the contemporary Bulgarian-born artist [SCIENTIFIC AMERICAN, June], contains an error in the discussion of the huge curtain planned for Rifle Gap in western Colorado. The project was not dropped for technical reasons.

Late last summer an attempt was made by Christo and his co-workers to put a curtain across Rifle Gap precisely as described in the review. Because of an unfortunate combination of errors, oversights and bad luck, the attempt was not successful, and the curtain was damaged beyond repair.

This summer Christo is having a new curtain made, additional work on the foundations completed, and the curtain is planned for erection early in August. Problems encountered in the first attempt have been analyzed and circumvented, and there is every expectation that a successful curtain-hanging will take place in Rifle Gap as planned.

JOHN H. THOMSON

Lynn, Mass.

Sirs:

We share the enthusiasm of ARPA for the commercial promise of computer networks ["Science and the Citizen," SCIENTIFIC AMERICAN, June]. In fact, we have been operating a commercial network that provides local call access to common information files for several years. This network currently covers 250 metropolitan areas in the U.S., Canada, Puerto Rico and, via satellite, much of western Europe. Further, both database-sharing and software-sharing are already well established commercial activities within this network.

The ARPANET approach of providing communications between existing geographically dispersed processors is one we regard as a significant interim step. Our studies had led us to the conclusion, however, that economies of much greater scale would flow from an approach in which the processors were centralized at a single site.

Confirmation of this conclusion has been obtained from commercial operation of our network, which embodies this approach. Additional benefits from processor centralization are being realized in the areas of security, accessibility, load-leveling across many time zones and the more effective use of hardware redundancy to improve reliability.

J. C. CASTLE

General Electric
Information Services Business Division
Bethesda, Md.

Sirs:

I should like to correct an error in my recent article "Cryptobiosis" [SCIENTIFIC AMERICAN, December, 1971]. In the drawing of the Cartesian diver (and also in the text) reference is made to a "seal" in the neck of the diver. Even students of elementary physics will recognize that if the diver is to be used to measure volume changes due to oxygen uptake, this seal cannot be a mechanical one, and indeed it is not. In Andrew Pigoń's original description of oil-filled Cartesian divers of high sensitivity, on which my own description of divers depended, he referred to a neck seal, meaning a seal composed of oil (*Bulletin de l'Académie Polonaise des Sciences*, Volume 1, page 65, 1953). The seal in the diver in my article should have been labeled "oil seal" and the text should have read something like "This was the diver; it was then weighted so that it had neutral buoyancy, and the neck was filled with oil."

JOHN H. CROWE

University of California
Davis

Scientific American, August, 1972; Vol. 227, No. 2. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the author's risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Harry T. Morris, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 25 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

Please enter change of address ☐
Please enter subscription ☐
☐ 1 year, $10 ☐ 2 years, $18 ☐ 3 years, $25

NAME

NEW ADDRESS

OLD ADDRESS

Merging theory into reality.

The dielectric lens finds a home on the range.

Here is the heart of today's newest range telemetry antenna. It's a dielectric lens, resulting from new materials and advanced circuit design developed at Raytheon.

Much smaller, lighter, and less complex than conventional phased arrays, the dielectric lens system provides off-axis, multiple target tracking at high rates, from a fixed or slowly moving mount. We're applying this same technology to airborne ECM, and it has exciting possibilities in many other applications.

This is just one part of our work in dielectric lens technology. And only one of many technologies where we're finding low-cost solutions for today's system requirements. Raytheon, the Systems Managers. Merging theory into reality wherever the need exists.

Raytheon Company, Lexington, Mass. 02173.

RAYTHEON

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1922: "In the early days of the petroleum industry the chief product was kerosene. The quantity of kerosene required to light the lamps of America and of such foreign customers as might be developed was the factor that determined the amount of petroleum refined. This condition has within the past two decades been reversed, however, and today it is the number of automotive vehicles in use and their requirements for gasoline—once a mere by-product of little value and less consequence—that govern the amount of petroleum refined, the manner in which the petroleum is split up into its fractions and the constant search for fresh sources of supply. The kerosene that in 1904 constituted practically 60 percent of the output is now less than one-eighth of that total. In 1904, when the automobile was just coming into its own, the automobile-juice represented an eighth of the refiner's business; today it is between a quarter and a third and we are worried because we cannot make it more. At the same time the heavy fuel-oil that shares the internal-combustion field with gasoline and competes with coal in the raising of steam has leaped from 16 to 55 percent of the refiner's total."

"The ether, or whatever may be the medium that carries the radio waves from the transmitter to the receiving set, must be policed. Ever since 1912 there have been in force certain laws and regulations, as well as a system of operators' and station licenses, for the use of the ether. These laws and regulations and licensing systems served admirably until the advent of radiophone broadcasting. Just as the rules of the road, such as they were, became hopelessly obsolete and inadequate when the automobile came into existence, so have the laws and regulations and licenses of 1912 become antiquated and obsolete in the face of the radiophone broadcasting activities throughout the nation. The Government, represented by the Department of Commerce, in charge of radio regulations, has been quick to realize the inadequacy of present radio laws, and has devoted no little time and effort, aided by the various groups interested in radio communication, toward the formulating of new laws and regulations."

"Long before industrial research laboratories were established in any country there existed a very thorough international cooperation in scientific research. The paths of scientific discovery were intentionally so blazed by the pioneers that others could follow. A wish to conceal natural laws has seldom developed among natural scientists. It would be difficult to find any considerable group of new phenomena that were purposely long concealed. It has always been bad form to bury a living truth. In the world as we fortunately find it most of those who study fundamental phenomena take pains quickly to add their contributions to the immense mass of carefully tested physical truth that constitutes the greatest inheritance of our civilization."

"Three members of the Everest expedition, Mallory, Somerville and Norton, on May 21 reached an altitude of 26,800 feet, the highest ever gained by man, and just 2,200 feet below the summit. To have got so far in a climb that was merely a kind of preliminary reconnaissance is a very fine achievement and seems to augur well for the success of the final effort."

SCIENTIFIC AMERICAN

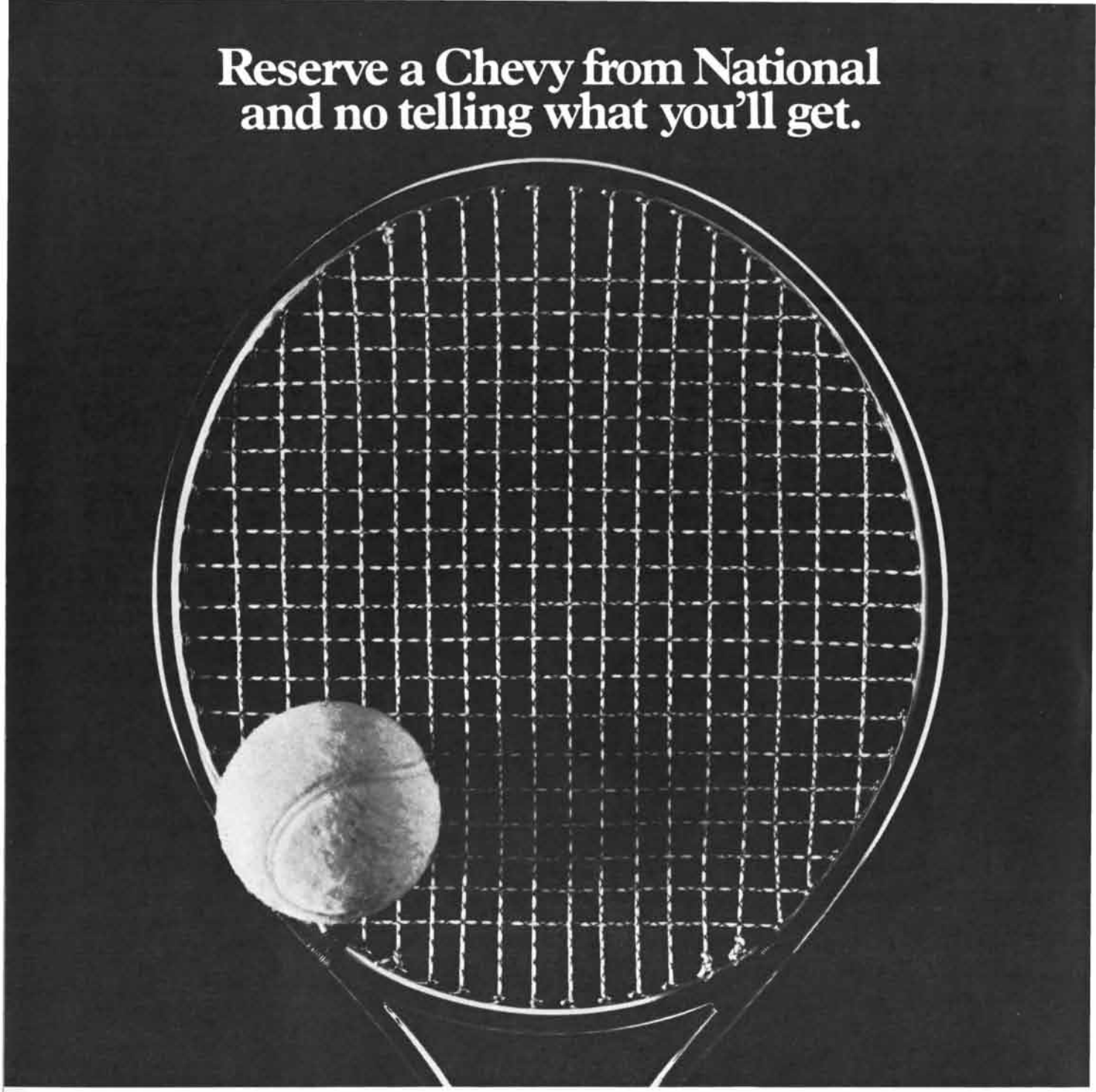
AUGUST, 1872: "The caisson on the New York shore for the Brooklyn suspension bridge is now filled in, and the erection of the stone tower will proceed as rapidly as possible. The tower on the Brooklyn side has reached the height of 105 feet above high water. The towers are to be 150 feet high. The wire cables will be 120 feet above the water. The span of the bridge will be 1,600 feet."

"Though we are compelled to think of space as being unbounded, there is no mental necessity that compels us to think of it as being either filled or empty; whether it is filled or empty must be decided by experiment and observation. Are the vast regions that surround the stars, and across which their light is propagated, absolutely empty? The reply of modern science is negative. The notion of the luminiferous ether must not be considered as a vague or fanciful conception on the part of scientific men. Of its reality most of them are as convinced as they are of the existence of the sun and the moon. The ether has definite mechanical properties. It is almost infinitely more attenuated than any known gas, but its properties are those of a solid rather than of a gas. It resembles jelly rather than air. It is the vehicle for star light, and without it stars could not be seen. If the ether has a boundary, masses of ponderable matter might be conceived to exist beyond it, but they could emit no light. Moreover, a body once heated there would continue forever heated. The loss of heat is simply the abstraction of molecular motion by the ether. Where this medium is absent no cooling could occur."

"An international congress is now in session in London, composed of representatives from all civilized countries, for the purpose of considering the questions of the repression and prevention of crime and the care of the criminal. Various questions relating to the subject of reform and punishment have been discussed at considerable length. Corporal punishment met with hearty condemnation from the American and Continental authorities. The question of the prevention of crime was also brought up, several members making reports of the labors of societies for that purpose. It was considered that the best mode of aiding discharged prisoners was by obtaining for them co-operative employment. The results of the deliberations of this congress of philanthropists cannot but be of the greatest importance. By this interchange of views the many and grievous faults of our present system of prisons and reformatories, which in a great measure are due to adherence to old and obsolete ideas, may be clearly seen and remedied, while valuable improvements and innovations will be suggested through the contrast of our methods of repression and prevention of crime with those adopted by other nations."

"The steep elevation of the lands immediately adjoining the city of San Francisco and the desirability of providing convenient access to these have induced some enterprising individuals to attempt the construction of an inclined railway. The incline will be 2,800 feet in length, traversed by cars drawn up by steel wire ropes and stationary engines. The cars are to be provided with clutches whereby the rope can be grasped or released at the will of the conductor. At the top of the incline the cars are delivered over to the horse railway."

Reserve a Chevy from National and no telling what you'll get.



Don't worry, you'll get the Chevy.

A nice new '72 Chevy. Or, if you'd rather, some other fine car.

But what you get afterwards, that's for you to say.

Because after you rent a car from us, we don't just smile and say thanks. We smile and say thank you with Green Stamps. Yep, good old S&H Green Stamps.*

And if you're in the habit of renting cars often, it doesn't take long for those stamps to really amount to something.

In fact, we've known our regular customers to redeem their stamps for things like a power lawnmower, an outboard motor, better golf clubs than they thought they could afford.

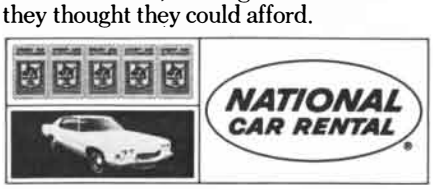
Take a look at an S&H catalogue and you'll get the idea.

One more thing deserves mentioning.

It doesn't matter to us whether you pay cash for your rental or charge it. The Green Stamps are yours.

Unless your wife sweet talks you out of them.

To reserve a car in the U.S. or elsewhere in the world call 800-328-4567 toll free.†



Take advantage of us.

National Car Rental System (in Canada it's Tilden Rent-A-Car). *Valid on U.S. rentals only except where prohibited by law. †In Minnesota call 612-944-1234 collect. © National Car Rental System, Inc. 1972

THE AUTHORS

DAVID E. COLE ("The Wankel Engine") is associate professor of mechanical engineering at the University of Michigan, where between 1960 and 1966 he obtained his bachelor's, master's and doctor's degrees. He writes that he has been interested in "things technical" since he was quite young, partly because his father was technically minded and partly because of summers that Cole spent on his grandfather's and uncle's farm. "Engineering," he notes, "is an important aspect of any agricultural business." (His father, Edward N. Cole, is now president of the General Motors Corporation.) Cole's interests apart from his work include trout fishing and home carpentry.

ECKHARD H. HESS (" 'Imprinting' in a Natural Laboratory") is professor of psychology at the University of Chicago. He is primarily an ethologist, having studied such diverse animals as the common fruit fly, fishes, chicks, ducks, geese, turkeys, rats, cats, guinea pigs, whales and man. His work with humans has included studies of changes in the aperture of the pupil of the eye resulting from attitudes, mental processes and emotional states; the most recent of his four previous articles in SCIENTIFIC AMERICAN was "Attitude and Pupil Size" in the April 1965 issue. Hess spends several months each spring at a research station he established in the saltwater marshes of the Eastern Shore of Maryland, where he carried out the work he describes. He received his Ph.D. from Johns Hopkins University in 1948.

RONALD BRESLOW ("The Nature of Aromatic Molecules") is S. L. Mitchill Professor of Chemistry at Columbia University. His degrees are from Harvard University: a bachelor's degree in chemistry, awarded *summa cum laude* in 1952, a master's degree in biochemistry in 1954 and a Ph.D. in organic chemistry in 1955. He joined the Columbia faculty in 1956. His wife, Esther Breslow, is on the faculty of the biochemistry department of the Cornell Medical School. Breslow writes that in addition to the work he describes his research interests include "work in biochemical model systems," involving "attempts to develop 'artificial enzymes' that will work in the way natural enzymes do, and also attempts to imitate some of the useful synthetic reactions that enzymes can perform. All these efforts have a common thread: the preparation of unknown molecules that are expected to have unusual and interesting new properties."

BART J. BOK ("The Birth of Stars") is professor of astronomy at the University of Arizona. Born in the Netherlands, he took his bachelor's degree at the State University of Leiden in 1926 and his Ph.D. at the State University of Groningen in 1932. From 1933 to 1957 he was at Harvard University. He then spent nine years as professor of astronomy at the Australian National University and director of the Mount Stromlo Observatory before going to the University of Arizona in 1966. Bok's work in the Southern Hemisphere included a period at Harvard's Boyden Station in South Africa, where he worked in 1950 and 1951.

BARBARA M. KORSCH and VIDA FRANCIS NEGRETE ("Doctor-Patient Communication") are at the Childrens Hospital of Los Angeles; Dr. Korsch is a physician and Mrs. Negrete is a public health nurse. Dr. Korsch, whose family left Germany in the Nazi period when she was 13, was graduated from Smith College and received her M.D. from the Johns Hopkins Medical School in 1944. After work in pediatrics and public health she became director of a large outpatient clinic at the Cornell University Medical Center, where she began the work on doctor-patient relations that she has continued in California. Her other area of professional interest is kidney transplantation; she is patient-care coordinator for a team of physicians and surgeons and associated health workers who are involved with kidney patients. "Aside from my professional life," she writes, "I am happily married to a pediatrician who is distinguished in the field of infectious diseases and virology. I have a charming son also." She adds that she enjoys cooking, particularly French dishes. Mrs. Negrete, who received her bachelor's degree at the University of California at Los Angeles in 1959 and her master's degree at Loma Linda University in 1963, is a research associate of Dr. Korsch's and nurse coordinator in the kidney program. She is married and the mother of an infant son.

F. G. HEATH ("Origins of the Binary Code") is professor of computer engineering at Heriot-Watt University in Edinburgh. After receiving his Ph.D. from the University of Manchester he worked in the electronics industry for several years, returning to Manchester to join the faculty of the Institute of Science and Technology at the university (he later became professor of digital processes there) and to serve as chief engineer of the Manchester engineering facilities of International Computers Limited. In 1970 and 1971 he was visiting professor in computing and information science at Case–Western Reserve University in Cleveland. Heath writes that he became interested in the history of digital codes some 20 years ago when he was told that a supposedly original coding system his team was using for a military guidance project had in fact been invented more than 80 years earlier by the French telegraphic engineer Emile Baudot.

JOHN D. PETTIGREW ("The Neurophysiology of Binocular Vision") is a postdoctoral fellow at the University of California at Berkeley. "I am a native-born Australian," he writes, "and grew up in the Blue Mountains west of Sydney. The mountains are in my blood and I find their allure as irresistible as that of neurons. When I am not in the lab trying to work out the intricacies of some cortical neuron's personality, I am finding my way through the forest, rock, scree and snow of some peak." Last summer he and his wife rode to Alaska on motorbikes for a mountaineering expedition. He notes that his favorite animal is the owl, "although I have a soft spot for Siamese cats, of which there are five in our household." Pettigrew's degrees are from the University of Sydney, the most recent ones being an M.B. and a B.S. from the medical school in 1969. He says that his ambition "is to help apply in Australia the ecological lessons that are painfully being learned here and in Japan."

IRA PASTAN ("Cyclic AMP") is chief of the Laboratory of Molecular Biology at the National Cancer Institute. He was graduated from Tufts College in 1953 and obtained his M.D. at the Tufts University Medical School in 1957. After internship and residency at the Grace–New Haven Hospital of the Yale University School of Medicine he joined the National Institutes of Health in 1959 as a clinical associate in the National Institute of Arthritis and Metabolic Diseases. He joined the National Cancer Institute in 1969 as head of the molecular biology section of the endocrinology branch, moving to his present position in 1970.

Established 1845 SCIENTIFIC AMERICAN August 1972 Volume 227 Number 2

The Wankel Engine

The demands of emission control and cost reduction are making this rotary power plant increasingly attractive for the U.S. automobile. Its proper use probably calls for a redesign of the entire vehicle

by David E. Cole

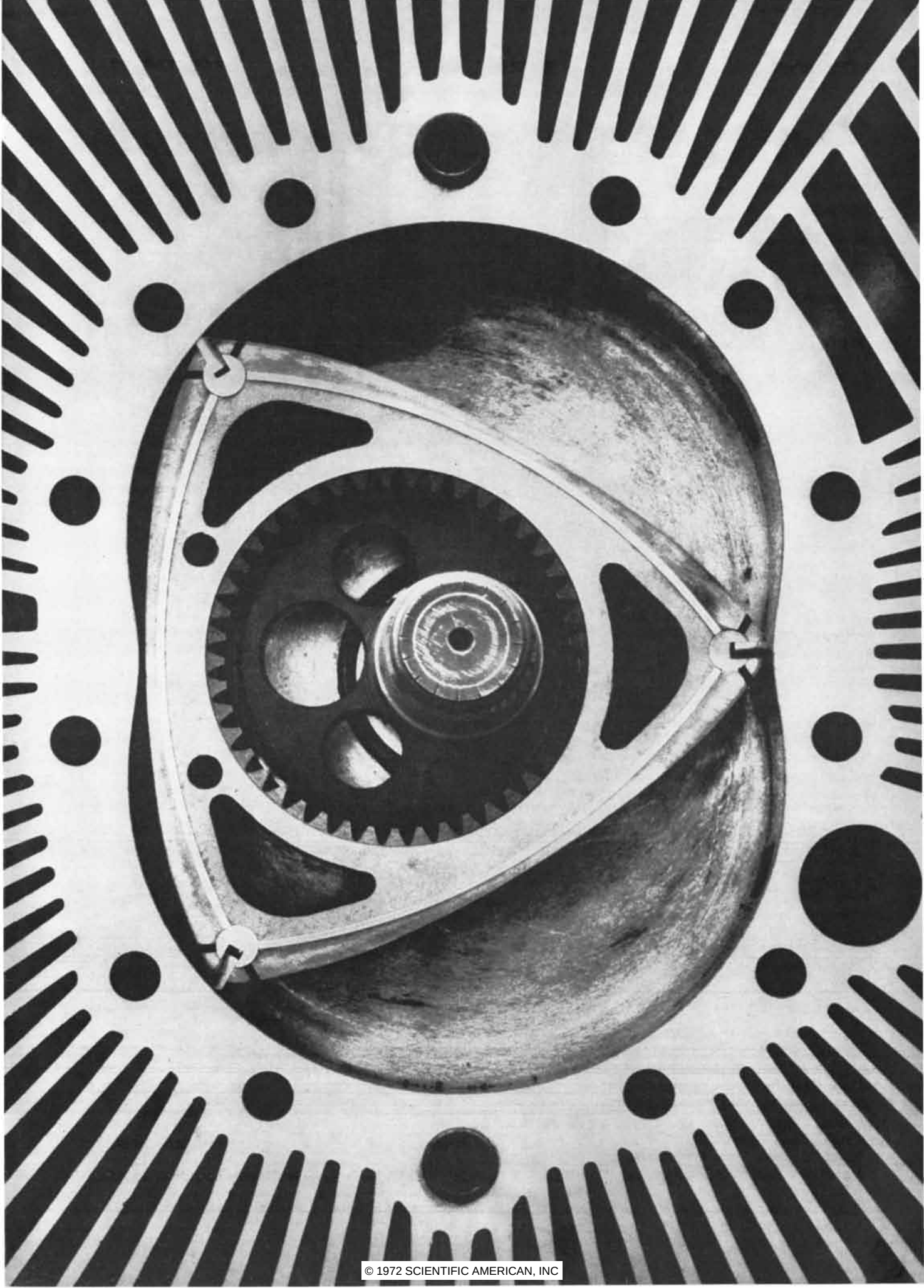
A technological revolution is gathering force in the U.S. automotive industry, stimulated by new Federal laws requiring a reduction in the pollutants contained in automobile exhaust gases. The industry is also under heavy pressure to reduce manufacturing costs in order to remain competitive with foreign automobile makers. For both reasons there are strong incentives to replace the piston engine whose four-stroke cycle was first demonstrated 96 years ago by Nikolaus August Otto. Numerous alternatives are being explored, including the gas turbine, electric propulsion and a variety of steam or vapor engines. Each of these, in comparison with the piston engine, offers certain emission advantages—and certain disadvantages. No prototype engine has passed the 1976 Federal emission standards as they are currently written. Emission control, however, is only one aspect of a complex challenge. In evaluating any power system one must take care to preserve a balanced outlook. For reasons that will become clear I shall argue that the most promising successor to the piston engine is the engine invented by Felix Wankel and known variously as the Wankel engine, the rotary engine, the rotating combustion engine or the R.C. engine.

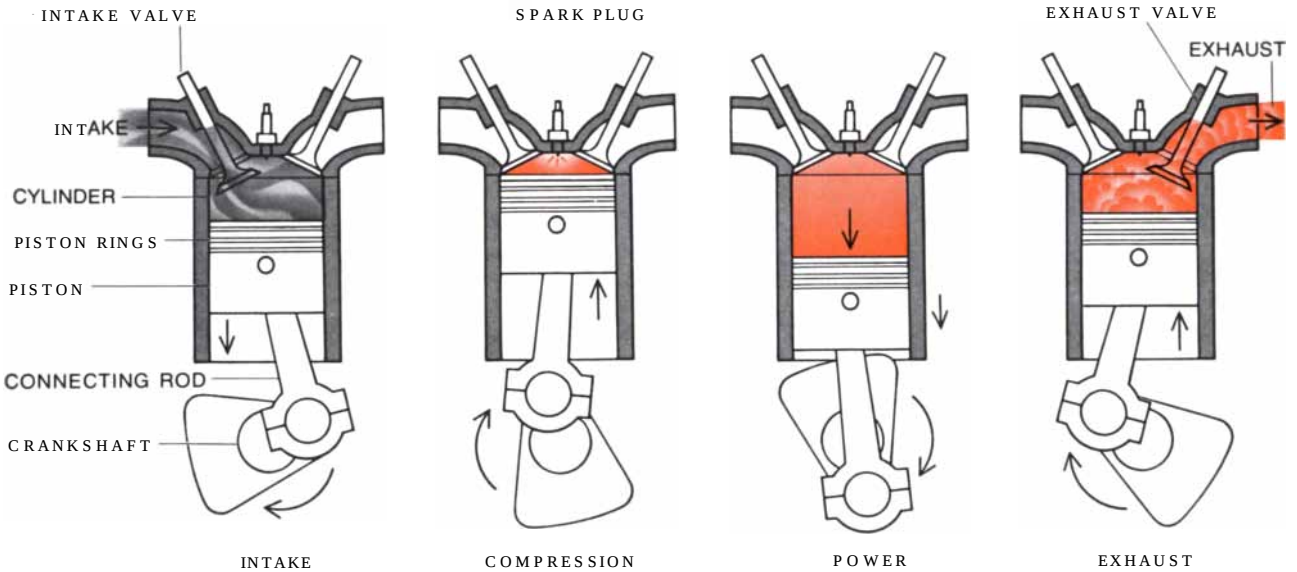
Unlike some of my colleagues in automotive engineering, I believe the internal-combustion engine can be controlled to satisfy future emission standards. I also believe the choice between alternative designs will be made not on the basis of emission control alone but rather on the basis of many technical and manufacturing considerations represented by the installation of an automotive power plant and all its appurtenances in an operating, competitively priced vehicle. Assuming an equivalent level of emission control, we can be sure that any competitor of the present-day automobile engine must perform as well as that engine in all respects—and better in some—before it can be regarded as a viable alternative.

In judging alternative power plants the automobile manufacturer will give weight to the following major considerations, in addition to exhaust emissions: thermal efficiency (meaning not only fuel economy but also respect for the conservation of resources), the use of critical materials, manufacturability, cost, size, weight, reliability, serviceability, noise level, safety and drivability. Failure to satisfy even one of these criteria is almost certain to disqualify an alternative design. With these stringent specifications in mind let us consider the Wankel engine.

In 1926 Wankel, a German engineer and inventor, began a systematic investigation of rotary engines. In the 1930's he established a research institute to examine the underlying physical principles more deeply, and he soon concluded that the successful development of a rotary engine awaited the solution of three problems. First, the sheer number of possible arrangements and cycles for rotary engines had to be reduced; inventors had been nearly overwhelmed by them. Second, someone had to develop a proper thermodynamic and gas cycle with adequate port areas and timing of events. Third, and not least, most rotary engines of interest could not use devices as simple as piston rings to seal off the combustion chamber; someone had to solve the problem of sealing in several planes, with special attention to sealing the corners of a chamber.

SNOWMOBILE POWER PLANT is an air-cooled Wankel engine built by Fichtel & Sachs AG of West Germany under license from the German automotive firm NSU. The faceplate of the engine has been removed to show the triangular rotor, which revolves eccentrically (and clockwise) in an epitrochoid housing. The radial wedges are cooling fins. If the engine were operating, the chamber at the lower right would contain a fresh charge of fuel and air. The smallest chamber, the one adjacent to the left-hand face of the rotor, would contain a fully compressed charge. At this point the firing of a spark plug would initiate the power stroke. The chamber at the upper right would contain spent gases that were escaping through an exhaust port. The engine, which develops 20 horsepower at 5,000 r.p.m., weighs only 56 pounds, or 2.8 pounds per horsepower. The compression ratio is eight to one.





CONVENTIONAL PISTON ENGINE employs the four-stroke cycle first used 94 years ago in an engine built by Nikolaus August Otto. During the intake stroke a mixture of fuel and air is drawn into the cylinder. When the compression stroke has reduced the mixture to about 10 percent of its original volume, the mixture is ignited by the firing of a spark plug. The expanding gases produce the power stroke. During the exhaust stroke the products of combustion are driven out through the exhaust port.

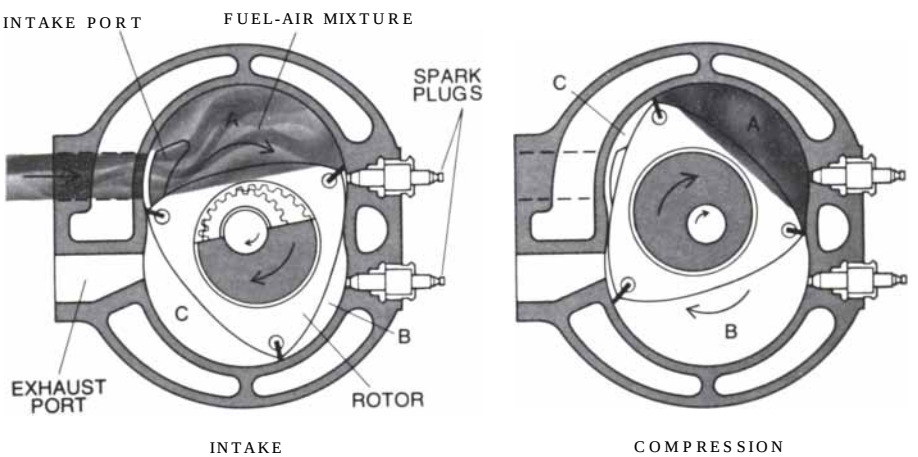
Wankel's initial mechanical concepts were complex and rather unpromising. Nevertheless, the German automotive firm NSU became interested in several of Wankel's ideas for rotary valve mechanisms and a compressor that might be applied in motorcycles. After a contract was signed between NSU and Wankel, Walter G. Froede of NSU became involved in the project. It was under his guidance that Wankel's compressor concept was first developed into an engine.

The first engine used two concentric rotors with parallel axes. A gear set connected the two rotors but did not transmit engine torque. Torque was delivered from the crankshaft of the inner rotor, which was also hollow to carry the air-fuel mixture. To reach the combustion chamber the mixture had to pass through intake ports in the inner rotor and cutouts in the side housing of the outer rotor. The complex design presented many problems and offered few advantages over the conventional piston engine.

With the assistance of Froede, Wankel kinematically inverted the original concept, thus creating an engine with a single rotor mounted in a stationary housing, both of unusual geometry. A prototype engine of this much more practical design was operated successfully in 1956. Its development received a big boost when the Curtiss-Wright Corporation bought the North American rights in 1958.

Under the direction of M. Bentele and Charles Jones, the Wankel engine was developed to the point where its potential as a mass-production power plant became increasingly evident. By 1960 Curtiss-Wright had demonstrated prototypes of its RC-2-60 series of Wankel engines that had two rotors on a common shaft and incorporated major advances in basic design. NSU combined innovations of its own with some of Curtiss-Wright's and introduced the first Wankel-powered vehicle, the NSU Spider, in 1965. Today both NSU and Toyo Kogyo of Japan are manufacturing "Wankel cars." The Mazda RX-2 and RX-3, which Toyo Kogyo has recently introduced into the U.S., are powered by two-rotor Wankel engines of 120 horsepower; it is an exciting vehicle with good driving characteristics, attributable largely to its engine.

In 1966 the Outboard Marine Corporation negotiated a U.S. license for the Wankel engine and four years later the General Motors Corporation became the first U.S. car maker to take a license. Today nearly every major engine builder in the world either holds a license or is negotiating for one. It has been strongly rumored (and denied) that one of the leading U.S. automobile companies will introduce a vehicle with a Wankel engine by 1974 or 1975. Not to be outdone, students at the University of Mich-



WANKEL, OR ROTARY, ENGINE has three small chambers of variable size, each of which undergoes a complete four-stroke Otto cycle in one revolution of the rotor. The five diagrams illustrate the sequence of events in the chamber labeled *A*. On the intake

igan have designed and are currently fabricating a two-passenger urban vehicle, powered by a Wankel engine developed by Outboard Marine for snowmobiles.

Although the Wankel engine is simple, its basic geometry is not easy to visualize. The inner working surface of the housing, within which the rotor turns, is called an epitrochoid. The term is given to the path described by a point within a circle rolling around another circle. The two-lobe epitrochoid of the Wankel is generated when the radius of the interior circle is twice the radius of the rolling circle [*see top illustration on next page*].

Since the apexes of the triangular rotor are always in contact with the housing, they create three separate variable volumes, each of which serves a key function. Unlike the piston in a conventional engine, which requires a connecting rod for transmitting power to the crankshaft, the Wankel's "rotary piston" runs directly on an eccentric shaft [*see bottom illustration on next page*]. Torque output is derived from this shaft. To maintain the rotor in proper orientation with the housing, an internal gear on the rotor is meshed with a gear fixed rigidly to the housing. Used strictly to maintain orientation, the gear does not extract power.

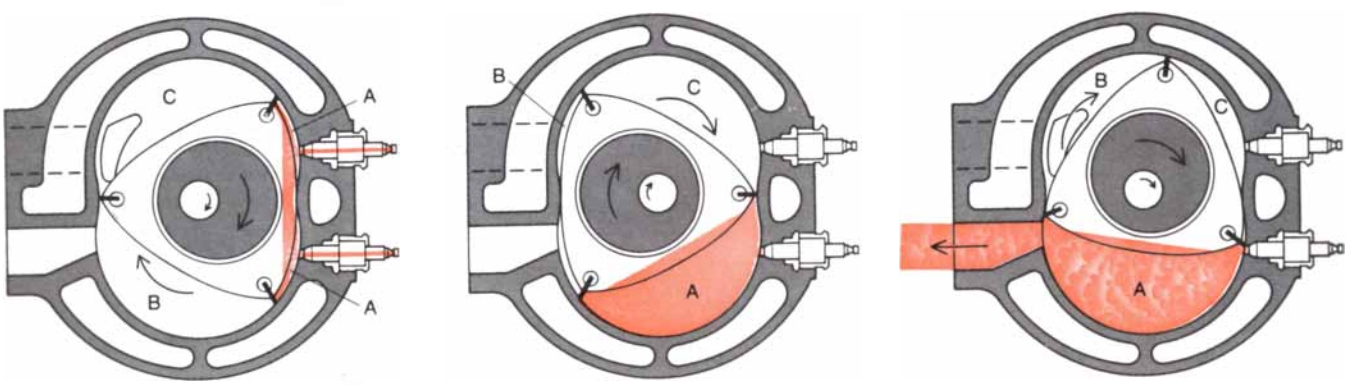
The Wankel also differs from the conventional automobile engine in having no valves. The fuel-air inlet port and the exhaust port are opened and closed at the appropriate time in the combustion cycle by the passage of the rotor itself. The sequence of events in the four-stroke Otto cycle of the conventional engine [*see top illustration on opposite page*] is probably familiar to most readers.

At the beginning of the four-stroke cycle the piston travels downward (first stroke), drawing a fresh charge of fuel and air through the open intake valve. Both the intake and the exhaust valves close and the piston ascends (second stroke), compressing the fuel charge. At the end of compression the charge is ignited by a spark, and the air-fuel mixture burns under conditions of nearly constant volume. The combustion increases the internal energy of the working fluid and therefore raises its temperature and pressure; as a result the piston is driven downward, producing the power stroke (third stroke). The energy is delivered to the crankshaft. The next time the piston travels upward (fourth stroke) the exhaust valve is open, so that most of the products of combustion are purged from the cylinder. The four strokes can be translated into a simple four-sided closed curve representing the thermodynamic cycle [*see illustration on page 20*].

The radically different operation of the Wankel can be more easily understood if one refers to the five-part illustration at the bottom of these two pages. Three chambers are labeled *A*, *B* and *C*; let us consider volume *A* alone. In Step 1 the intake port is uncovered and volume *A* is expanding, drawing a fresh air-fuel charge into the engine. In Step 2 the trailing apex seal has isolated volume *A* from the intake port and the air-fuel charge is compressed as volume *A* steadily decreases with rotation of the rotor, reaching the minimum shown in Step 3. At this point a spark plug ignites the charge and chemical energy is converted into the internal energy of the working fluid. Step 4 is the expansion, or power, stroke, in which high-pressure gases drive against the eccentric rotor. In Step 5 the leading apex seal on the rotor face forming volume *A* passes and uncovers the exhaust port, allowing combustion gases to be purged as the rotation of the rotor drastically shrinks volume *A*. While this sequence of events is taking place, the other two faces of the rotor are defining volumes *B* and *C*, which are going through exactly the same sequence (slightly offset, of course, in time). Thus three different events—one at each rotor face—are always occurring simultaneously.

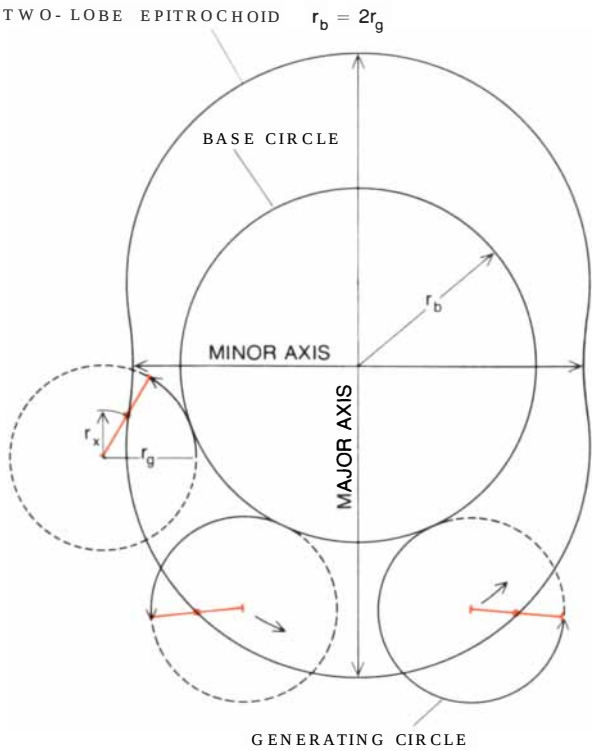
The Wankel's crankshaft makes three rotations for every one rotation of the triangular piston, which means that the engine delivers one power stroke for each full rotation of the crankshaft. This is exactly twice the power-stroke frequency in the conventional four-stroke engine. In other words, the Wankel engine uses its displacement volume twice as often as the four-stroke power plant does. One immediate consequence is that for equivalent horsepower a Wankel engine is only about half the size and weight of a conventional engine. Moreover, since the Wankel has no valves, it does not need a camshaft, valve lifters and so on, and it requires no more than two spark plugs per rotor. A typical V-8 engine of 195 horsepower has 388 moving parts, but one American-built two-rotor Wankel of 185 horsepower has only 154 moving parts.

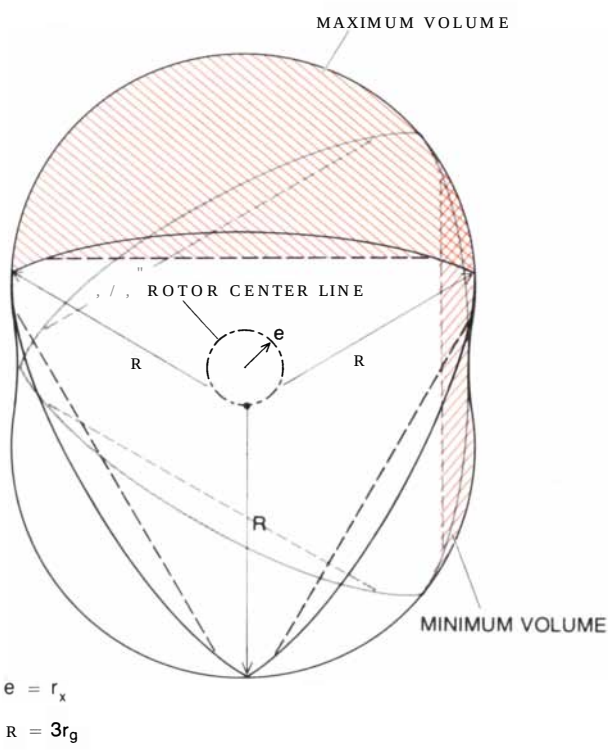
The Wankel engine should prove more reliable and serviceable than its conventional equivalent if only because of its comparative simplicity. It has been estimated that if and when Wankel engines are put into mass production, they



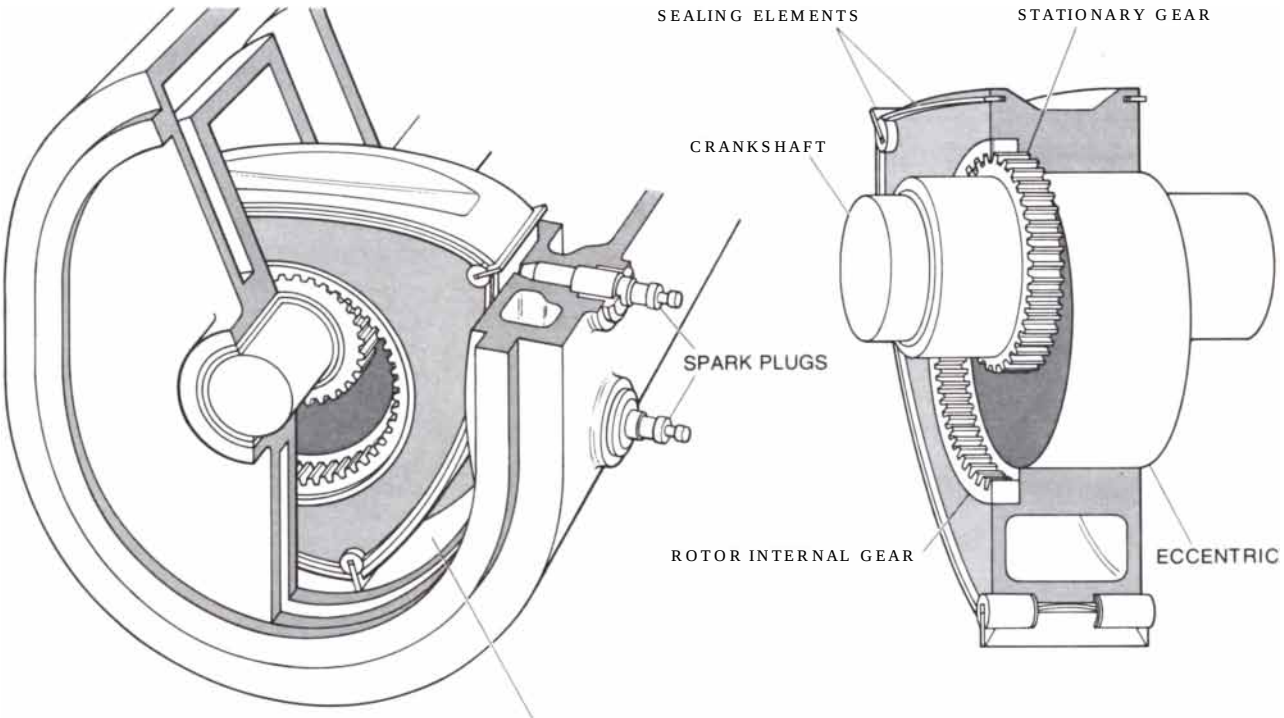
IGNITION EXPANSION EXHAUST

stroke a fresh charge of fuel and air enters the chamber through the intake port, which is always open. As the rotor turns clockwise the charge is sealed off and compression begins. When the charge is fully compressed, it is ignited by the firing of a spark plug. The expanding combustion gases drive against the rotor until the exhaust port is uncovered, allowing the spent gases to escape.





WANKEL'S EPITROCHOID HOUSING is a geometric figure (*left*) traced by a point in a generating circle as it rolls around a base circle. In the standard, or two-lobe, configuration the radius of the base circle, r_b, is twice the radius of the generating circle, r_g. The same curve can also be generated by the apexes of a triangular rotor (*right*) of radius R whose center is displaced from its center of rotation by some value e (the eccentricity). The curve generated by the apexes of the rotor coincides with the curve generated by a generating circle when the rotor's radius R equals $3r_g$ and when e equals r_x, which designates the location of the point on the generating circle that is used to trace out the epitrochoid. The ratio R/e defines the theoretical compression ratio of the Wankel engine.



WANKEL'S CRANKSHAFT, which transmits the engine's torque, runs in sliding contact with the triangular rotor, as shown in the sectional view at the right. The gears do not transmit torque. The internal gear on the rotor meshes with a stationary gear that is rigidly attached to the housing, thus keeping the rotor and housing in proper orientation. The crankshaft makes three rotations for each rotation of the rotor. The motion of the rotor can be balanced completely with rotating counterweights attached to the crankshaft.

should cost only about half as much per horsepower as a conventional engine. That is a quite reasonable expectation, since automobile manufacturers have learned that given comparable costs for raw materials, the manufacturing cost of an automobile component is determined largely by its weight.

Unlike the piston engine, the Wankel is almost completely vibrationless. Since all motion in the Wankel is rotary, all unbalanced forces are simple harmonic and any imbalance can be fully canceled with rotating counterweights. The piston engine, on the other hand, has reciprocating imbalance with higher-order harmonics that cannot be dealt with in this way; at best effective balance can be approached only in certain multicylinder configurations where the unbalanced forces of one cylinder are equal and opposite to those of another. Such balancing is why a V-8 engine runs more smoothly than a four-cylinder in-line engine. The Wankel's relatively open flow path, unencumbered by restrictive valves, allows it to "breathe" far more easily than its piston-engine counterpart. One can grasp the significance of this advantage by remembering that any internal-combustion engine must function through much of its operating cycle as an efficient pump of the fuel mixture and the exhaust.

Torque in the Wankel is developed over some two-thirds of each crankshaft revolution as compared with only one-fourth of a revolution in the four-stroke engine. This dispersed application of torque adds to the basic smoothness of the Wankel. A single-rotor Wankel is as smooth as a three-cylinder piston engine. Most Wankels now being built for automotive use have two rotors.

The fuel-octane requirement of Wankel engines is lower than that of piston engines of the same compression ratio. Moreover, unleaded fuel causes no trouble. In the piston engine the lead in leaded fuel acts as a lubricant between exhaust valves and their seats. In the absence of lead the seats would wear rapidly unless specially treated (for example by induction hardening) or fabricated from alloys more expensive than those in common use today. The Mazda RX-2 runs well with unleaded fuel of between 80 and 85 research octane number. Typical 1972 automobile engines require fuel with an octane number of 91+.

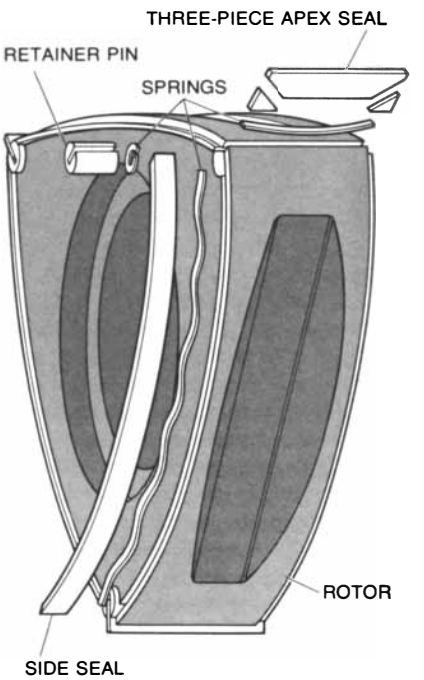
It is generally conceded that Wankel engines are quieter than comparable piston engines. The lower noise level is due partly to greater smoothness of operation and partly to the absence of parts that "ring." Valve clatter is of course eliminated along with the valves. The Wankel is also easier to muffle than the kind of two-stroke piston engine typically used in outboard motors, snowmobiles and chain saws. Exhaust noise from these engines is difficult to muffle because effective systems increase back pressure, and even a slight pressure buildup degrades the performance of the engine considerably. Because of its simplicity and low number of moving parts the Wankel operates with less friction than a conventional engine. The reduced friction, together with better breathing and absence of valve train, enables the Wankel to operate at higher speeds. The lower friction alone means decreased parasitic losses and can be translated directly into an increase in fuel economy.

Perhaps the most far-reaching economic significance of the Wankel lies outside the engine itself in the potential it offers for the redesign of the automobile chassis. The Wankel's comparatively small size may well provide the automobile manufacturer with one of the most dramatic cost-control opportunities of the century. A major dilemma—perhaps *the* major dilemma—of automobile manufacturers is how to reduce costs without reducing the value to the consumer. If the domestic automobile industry fails to resolve this dilemma by innovation, we can be sure the overseas industry will not be so timid.

Let us consider how the Wankel engine could contribute to the redesign of a typical American compact, such as the General Motors Vega, the Ford Pinto or the American Motors Gremlin. Not only would the rotary engine be lighter by half than the present engine; it also would allow the overall length of the car to be shortened by about a foot. Typically when the weight of a piston engine is reduced by 100 pounds, the easing of structural requirements allows the weight of the chassis to be reduced by at least 25 pounds. The weight reduction provided by the Wankel would yield this rule-of-thumb weight saving in addition to the weight saving obtained by shortening the chassis. The package immediately surrounding the engine contributes nothing to passenger comfort or safety; indeed, it is a liability. The rugged piston-engine block allows little crushing of the front structure; it actually tends to prevent the structure from absorbing the energy of impact in a collision.

With the adoption of the Wankel a front-engine, front-drive power train could become standard in the American automobile. Because the piston engine is so bulky a front drive is now found in only two of the more expensive American cars. With the much smaller Wankel a front-drive vehicle has compelling advantages [*see illustration on page 21*]. By eliminating the conventional transmission and drive-shaft "hump," designers could provide significantly more room for passengers. Additional luggage space could be gained in the trunk, and at the same time the fuel tank could be shifted to a less vulnerable position in front of the rear axle. The overall space saving would be particularly significant in a small car, where the drive train now takes up a disproportionate share of the room.

Aerodynamic losses could be markedly reduced below the present values by fitting the small Wankel into a low-drag package. Rolling resistance would also be reduced by the reduction in weight. (Rolling resistance on hard pavement is due primarily to hysteresis losses in the tires, and is thus directly proportional to the weight of the vehicle.) Considered

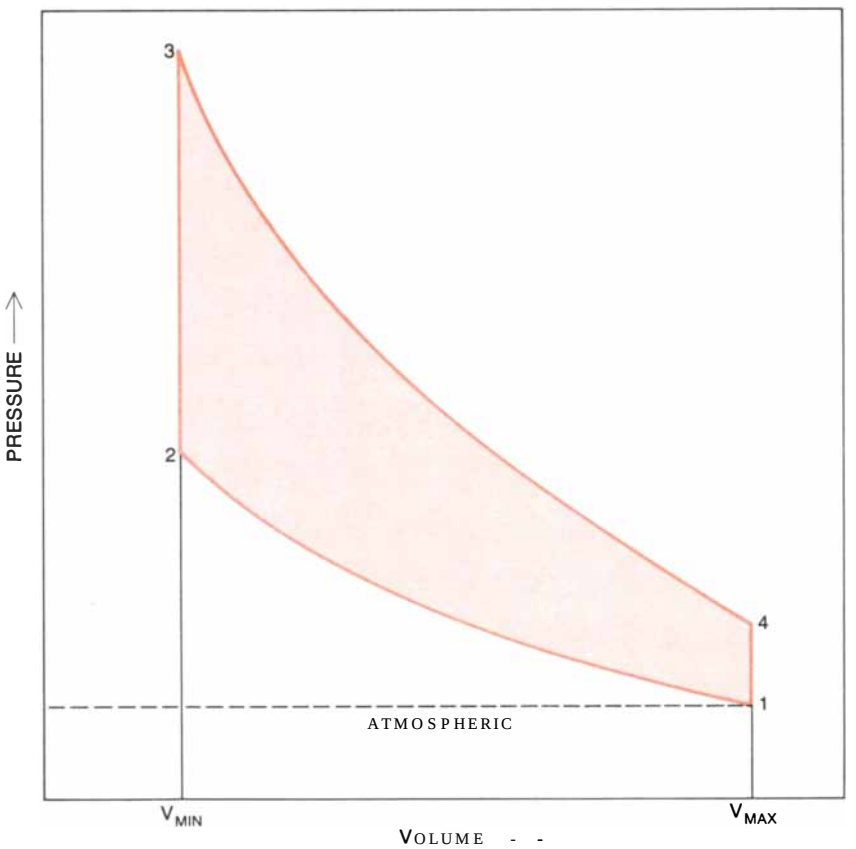


THREE-PIECE APEX SEAL is used on recent versions of the Curtiss-Wright rotating-combustion, or Wankel, engine. Two small triangular-shaped elements float freely against the main blade and provide effective seals at the corner of the chamber. Seals are pressed against the surface of the chamber by springs in combination with gas pressure that is bled into the region behind the seals. Retainer pin, a key element in sealing system, has proved to be an effective interface between the side seals and the apex seals.

together these potential advantages suggest that a Wankel-powered car would provide better performance and fuel economy than a piston-powered car of the same horsepower. More important, the level of exhaust emissions would be decreased because less power would be required for the performance of a given task.

The low-hood profile and better front visibility made possible by the Wankel engine should be one of several benefits for the driver. The smaller vehicle size and lower weight should bring improved handling and maneuverability. One little-understood and generally misinterpreted factor in vehicle control involves the concepts of understeer, neutral steer and oversteer. Stated simply, a vehicle with understeer requires the driver to turn the steering wheel at a greater and greater angle toward the center of the turn as speed is increased on a curve of constant radius. Under the same conditions a vehicle with neutral steer will negotiate the curve at all speeds with the same position of the steering wheel. In a vehicle with oversteer the driver must turn the steering wheel away from the center of the turn as speed increases. Of the three characteristics understeering is the most desirable and is normally exhibited by conventional rear-drive cars. Unfortunately when the driver is forced to perform emergency maneuvers that include sharp acceleration or deceleration, the understeer characteristic deteriorates markedly. Front-drive vehicles tend to respond to such emergency maneuvers in the opposite, hence desirable, way. Moreover, on slippery surfaces front-drive vehicles have much superior traction.

A decrease in vehicle weight made possible by adoption of the Wankel engine would dramatically increase the life of tires and brakes. The size of the radiator and the drive train (to name just two components) could be reduced, again for the simple reason that a Wankel automobile would require less horsepower for a given level of performance. The integrity of the overall structure could be strengthened, and thanks to reduced mass one could expect a substantial improvement in the vehicle's resistance to impact damage at low speeds. With a Wankel engine it should be possible to provide as much passenger and trunk space in a vehicle of intermediate size as one now finds in Detroit cars of standard size (for example the General Motors Impala, the Ford LTD or the Chrysler Fury). Overall the Wankel-powered car would probably weigh from 600 to 1,000 pounds less than the standard car, equivalent to a saving of $600 to $1,000, on the basis that present cars retail at about $1 per pound. The potential saving should more than pay for the safety features and emission-control systems that will be required in the future.

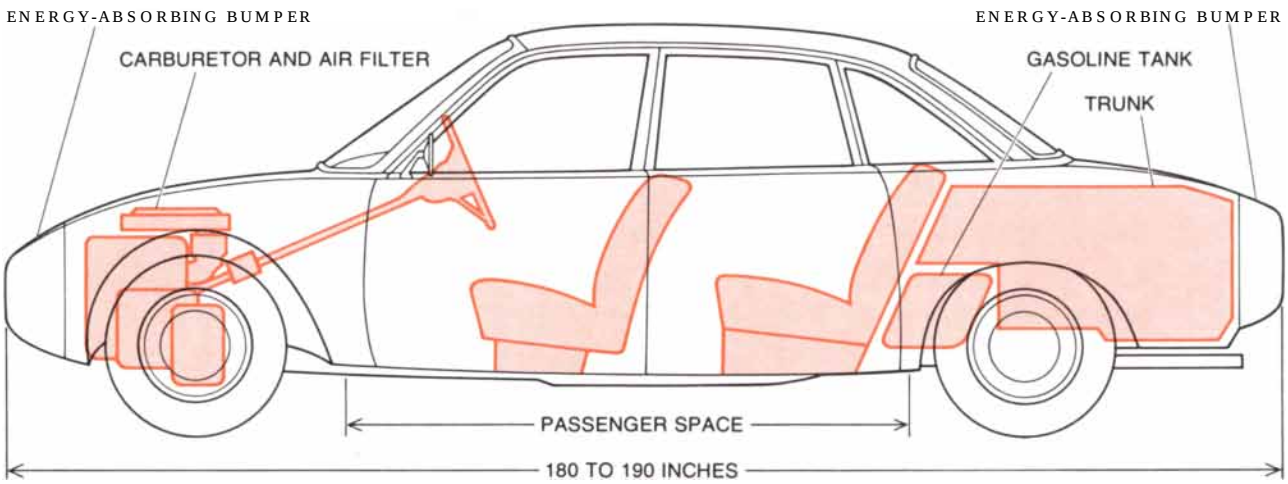


IDEAL THERMODYNAMIC CYCLE for the spark-ignited internal-combustion engine, whether conventional or rotary, is called the Otto cycle. A mixture of fuel and air is compressed reversibly and adiabatically from state *1* to state *2*. Reversible processes are "perfect"; "adiabatic" means without heat loss or gain. At point *2* the mixture is ignited, converting chemical energy into internal energy without change in volume so that the pressure rises to point *3*. The power, or expansion, stroke carries the system reversibly and adiabatically from state *3* to state *4*. With the opening of the exhaust port the system returns to state *1*. The ideal work output of the cycle is equal to the area enclosed by the curve.

So much for the advantages of the Wankel. What are its controversial features? One of the biggest is the mere fact that the engine is new. Even engineers and businessmen who are particularly open to innovation approach major expenditures in a conservative mood. Nevertheless, the fact that the Wankel is basically the familiar internal-combustion engine, and so should not intimidate the average repairman, ought to make its acceptance fairly rapid. In any case the demonstrated marketing capabilities of American industry can help to circumvent the difficulties caused by newness. Considering too that American merchandising thrives on introducing something new (however deserving of the adjective), fear of the Wankel's novelty may prove groundless.

From the beginning Wankel engines faced three major problems that persist to some degree in current designs. They are the seal between the rotor and its epitrochoid housing, fuel economy and exhaust emissions. There have been other problems as well, but they have responded favorably to engineering development.

The seal problem has proved to be particularly refractory. The most critical leakage site is at the three apexes of the triangular rotor. Although the side seals are equally important, they present a far simpler engineering problem. In existing designs each apex seal consists of a single linear structure that may have several components [*see illustration on preceding page*]. It is clear that the single apex seal must be cleverly designed to work as well as the three or more rings mounted on each piston of a conventional engine, which prevent the high-pressure gases in the combustion chamber from leaking into the crank-



HYPOTHETICAL WANKEL AUTOMOBILE with front-wheel drive would provide as much passenger and luggage space as is now found in full-sized American cars but would be as much as 30 inches shorter and some 600 to 1,000 pounds lighter than the conventional car. The Wankel engine, transmission and differential would all be mounted transversely. The front-wheel drive would eliminate the "hump" now found in the floor of most rear-wheel-drive vehicles. The downward-sloping hood, made possible by the small size of the Wankel engine, would improve the driver's view of the road. The gas tank is moved forward for safety.

case. Some automotive engineers have gone on record as stating that the problem of sealing the Wankel rotor is insurmountable. Others prefer to view the problem as a challenge and not as an attempted violation of a law of nature. The apex seal in current rotary engines (such as the Mazda, which uses at the apexes a single piece of carbon impregnated with an aluminum alloy) cannot yet match piston rings in durability. It has been reported, however, that newly developed sealing materials can provide more than 100,000 miles of service.

Poor fuel economy has been a common complaint, and undoubtedly a valid one, of the owners of the early Wankel-powered cars. There are two probable reasons for high fuel consumption: fuel combustion in the Wankel tends to be slower and hence less efficient than it is in the conventional engine, and some of the fuel-air mixture leaks past the sealing elements. In the latest designs combustion is accelerated by using two spark plugs in each combustion chamber. This stratagem, together with improved seals, has reportedly raised the fuel efficiency of current prototype Wankel engines to the level of piston engines of equivalent power. When the prototype engines are installed in well-designed, low-weight vehicles, they should give fuel mileage as much as 10 to 15 percent better than a piston engine in a conventional automobile.

In the Wankel the lubrication of the internal engine components presents no special problem except for the apex seals. Unlike the cylinder walls in a piston engine, the surface between the apex seal and the housing is never exposed to a bath of oil. In the small Wankel engines used in snowmobiles and the like, oil is mixed with the fuel. In engines of larger size, however, an oil-injection system with a metering pump is required. Such an injection system need not be expensive, and an oil-consumption rate of one quart per 1,000 or 1,500 miles seems readily attainable. In future designs oil could also be metered to the critical contact region between the apex seals and the housing by arranging for controlled leakage of oil past the seals. Because the Wankel burns oil at a controlled rate while exposing the remaining oil to a less stringent environment, it could put an end to the 6,000-mile oil change. I can envision the owner's manual of a 1980 Wankel car suggesting that oil be changed only every 50,000 miles, with oil to be added only as required. It is even possible that the engine, the transmission and the differential could use one oil reservoir in common.

One early problem encountered with the Wankel was sustained uneven heating of the housing. For example, the section near the intake port remains relatively cool, since it is exposed to the incoming air-fuel mixture, whereas the region near the spark plug is constantly heated by the combustion of the fuel. Therefore ways must be found to cool the combustion section of the housing and to counteract the stresses caused by the large temperature gradient between one section of the housing and another. Both problems proved to be soluble even in some of the earliest designs.

Of the three major problems I have mentioned, the important matter of exhaust emissions has had the most interesting history. When the Wankel first appeared in the late 1950's, the pollution due to automobile exhaust gases was just being generally recognized. Unburned hydrocarbons were the first major pollutant to come under scrutiny. The Wankel engine seemed inherently to release more unburned hydrocarbons than piston engines (largely because of the sealing problem), and so the concept of the rotary engine was summarily rejected by the major manufacturers. Taking a short-term approach, they failed to recognize that even the piston engine might be unable to meet future restrictions on the emission of unburned hydrocarbons and, what is more to the point, that additional pollutants, beginning with carbon monoxide and nitrogen oxides, would eventually be subjected to strict regulation. (The principal oxide of nitrogen emitted by the internal-combustion engine is NO; NO_2 is produced in small amounts. The mixture is commonly designated NO_x.)

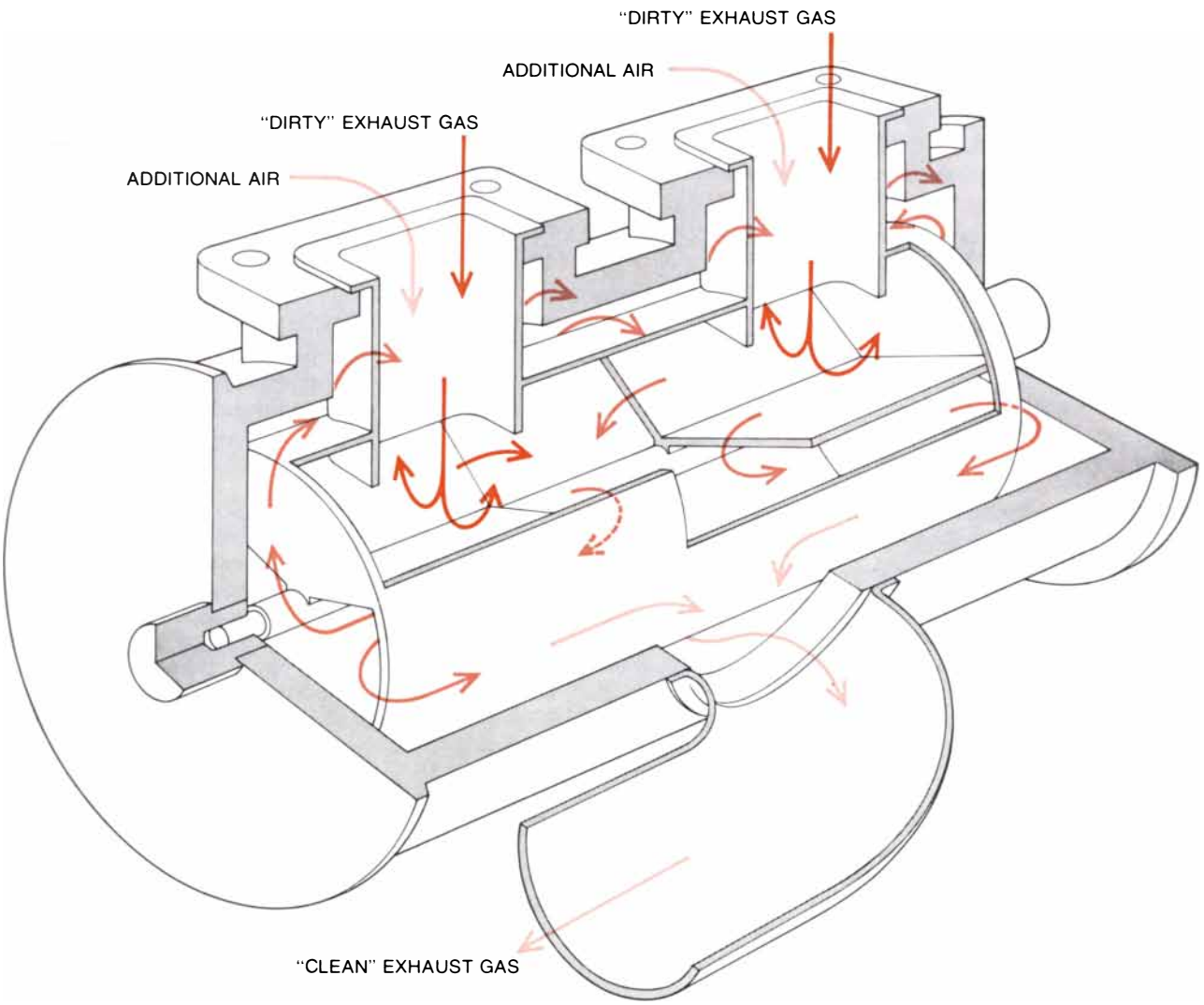
With this background in mind, let us take note of the Federal emission standards that have been set for 1975 and 1976. Based on an emission level measured in grams per mile for a simulated driving cycle, the 1975 standards are .41 gram for unburned hydrocarbons and 3.4 grams for carbon monoxide. For 1976 the NO_x limit is set at .4 gram per mile. These standards represent a reduction of 90 percent below the levels stipulated for 1971–1972. Achieving the 1975–1976 goals promises to be so difficult that some spokesmen for the auto-

motive industry are already suggesting it cannot be done.

Would the Wankel make the job any easier? In recent Wankel engines the quantity of unburned hydrocarbons has been reduced markedly by improvements in sealing and in preparation of air-fuel mixtures. The production of carbon monoxide is essentially a function of the fuel-air ratio, so that for any given set of operating conditions the Wankel and the piston engine can be considered comparable.

Nitrogen oxides result from the reaction of atmospheric nitrogen with oxygen. Their combination is enhanced by three factors: high maximum temperatures, long residence time at elevated temperatures and slightly lean fuel-air mixtures, which maximize the availability of nitrogen and oxygen without significantly lowering the combustion temperature. Emission data from existing Wankel engines of several designs show that they release only about half as much NO_x as piston engines operated under the same conditions. This is an important finding, because NO_x is proving to be the most difficult automotive pollutant to control. The low emission levels of the Wankel can be attributed to a number of factors, including the high surface-to-volume ratio of the combustion chamber and slower burning (both of which reduce the maximum cycle temperature) and perhaps the rapid cooling of the hot gases when they expand through the necked portion of the epitrochoid chamber near its top dead-center position.

If either Wankel or piston engines are to meet the 1975–1976 emission standards for unburned hydrocarbons and carbon monoxide, it seems clear that they will require in their exhaust line either a thermal reactor or a catalytic reactor (or a combination of the two). A thermal reactor is simply a fairly large, well-insulated exhaust manifold designed to bring about more complete oxidation of the residual hydrocarbons and carbon monoxide [*see illustration below*]. Catalytic reactors can be of either the oxidizing or the reducing type. (Since the reducing reactor is specifically for reducing NO_x to free nitrogen and oxygen it must be combined with the oxidizing reactor.) Both thermal and catalytic reactors work better with the Wankel than with the piston engine be-



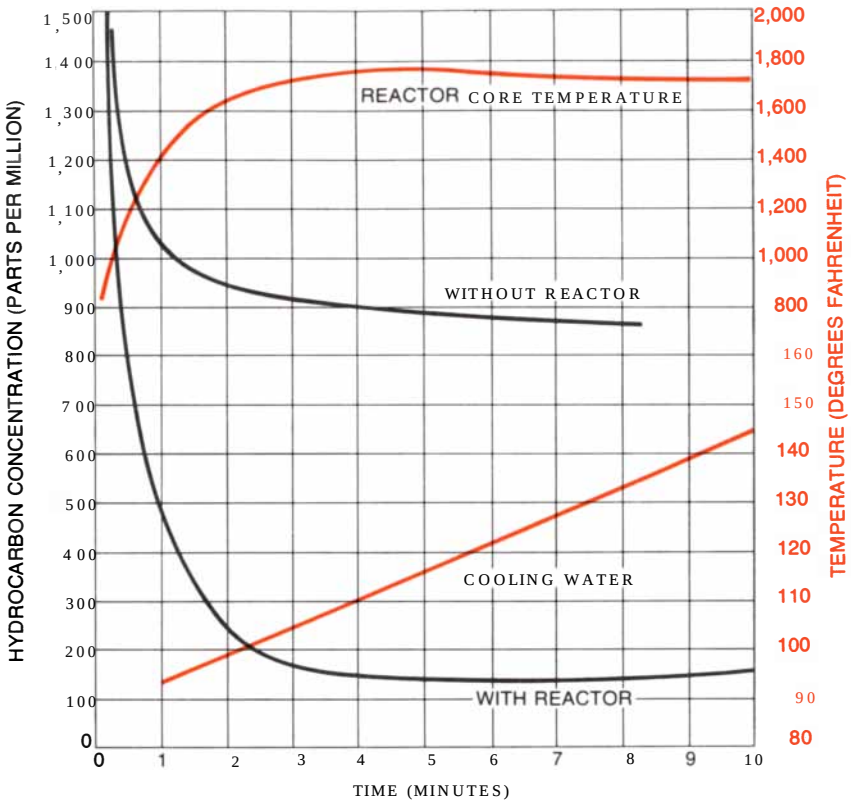
THERMAL REACTOR is an "after-engine" device designed to reduce the amount of unburned hydrocarbons and carbon monoxide present in exhaust gases. The reactor, which is well insulated to minimize heat losses, also acts as an exhaust manifold. If the temperature in the reactor is high enough, if air (oxygen) is plentiful and if the air and the "dirty" exhaust gas are well mixed, most of the hydrocarbons and carbon monoxide are removed by burning. Gases are directed through several passages to achieve good mixing. The reactor depicted here was used in a Wankel engine emission study at the University of Michigan sponsored by Curtiss-Wright.

cause of the Wankel's higher exhaust-gas temperature and faster exhaust warm-up.

A high exhaust-gas temperature promotes a rapid oxidation reaction if excess oxygen is present; a rapid warm-up shortens the interval between the time the engine is started and the time the exhaust-treatment device becomes effective. The Wankel's average exhaust temperature is high for several reasons. First, the exhaust gases pass through the exhaust port twice as often as they do in a typical four-cycle engine. Second, the exhaust passage remains hot because there is no exhaust valve that must be protected by nearby water jacketing. Third, the Wankel's slightly lower flame speed tends to raise the temperature near the end of the expansion stroke. Fourth, in two-rotor designs the exhaust ports are so close to each other that one can have a small exhaust collector with a low surface-to-volume ratio, thereby reducing heat losses. The Wankel's rapid exhaust warm-up follows from the same factors that give rise to high average temperatures. Furthermore, the Wankel's low weight per horsepower means less metal to absorb heat when the engine is started.

A thermal reactor alone is capable of promoting the effective oxidation of hydrocarbons and carbon monoxide, provided that the exhaust gases are held in the reactor long enough, that oxygen is mixed with the gases well enough and that exhaust temperatures are high enough. The importance of these three factors was proved by early research at the University of Michigan [*see illustration on this page*]. A thermal reactor is used in current Wankel-powered engines being manufactured by NSU and Toyo Kogyo for sale in the U.S. Emissions of NO_x are not significantly influenced by a thermal reactor.

The oxidizing catalytic reactor seems to be at least as effective as the thermal reactor for removing carbon monoxide and hydrocarbons, and it has the added advantage of being able to handle minor engine maladjustments without serious loss of performance. A catalytic reactor works as well with a Wankel engine as with a piston engine provided that the catalyst bed is protected from overheating. For cutting down NO_x emissions a special reducing catalyst would be highly desirable. Unfortunately a reducing atmosphere can be achieved only through the careful control of the mixture ratio, and a noxious product (ammonia) is easily produced. The problem is currently under investigation.



THERMAL REACTOR ATTACHED TO WANKEL warms up rapidly even at simulated idling speed, reducing hydrocarbon emissions to less than 200 parts per million compared with 900 parts per million in the absence of a reactor. Note that the reactor core heats up much more quickly than the Wankel engine itself, as evidenced by the fairly slow rise in the temperature of the engine's cooling water. When a similar reactor is attached to a conventional automobile engine, the reactor core warms up more slowly, with the result that about four minutes are needed to reach minimum emission of unburned hydrocarbons compared with only two minutes with a Wankel engine. The data plotted here were obtained in the author's University of Michigan laboratory with the very first reactor design that was tested.

By suitable engine design and adjustment of the operating variables it may be possible to control the Wankel's output of NO_x without exotic devices for cleaning up the exhaust and without the power-stealing recirculation of exhaust gases required in cars now being sold in California. One promising direction offered by the Wankel involves "stratification" of the fuel-air charge, a condition in which the mixture is not homogeneous throughout the volume of the combustion chamber. Stratification of the charge can readily be accomplished by injecting the fuel into a chamber already filled with air, but even without fuel injection existing Wankel engines exhibit considerable stratification. A rich mixture in the vicinity of the top dead-center position of the rotor will substantially decrease the peak cycle temperature and therefore decrease the amount of NO_x. A lean mixture in the leading section of the chamber will enhance the oxidation of carbon monoxide and hydrocarbons and simultaneously provide a built-in air-injection system for an exhaust reactor.

In summary, although the exhaust from the Wankel engine contains slightly more unburned hydrocarbons than the exhaust from a piston engine, it contains less NO_x. Exhaust clean-up devices respond more favorably to the Wankel's operating conditions. This factor, together with the potential for reducing the size and weight of a vehicle built around the Wankel, makes the engine a strong candidate as a "clean" power plant.

Prophecy is always risky, but it seems clear that the Wankel engine (or rotary engine, as it will probably be known) will loom large in the future of the engine industry. Wankel engines small enough for lawn mowers and chain saws, and others larger than any needed for automobiles, have been designed and are currently being intensively tested. Future generations may well rank Felix Wankel with Nikolaus August Otto as the progenitor of a major technological revolution.

"Imprinting" in a Natural Laboratory

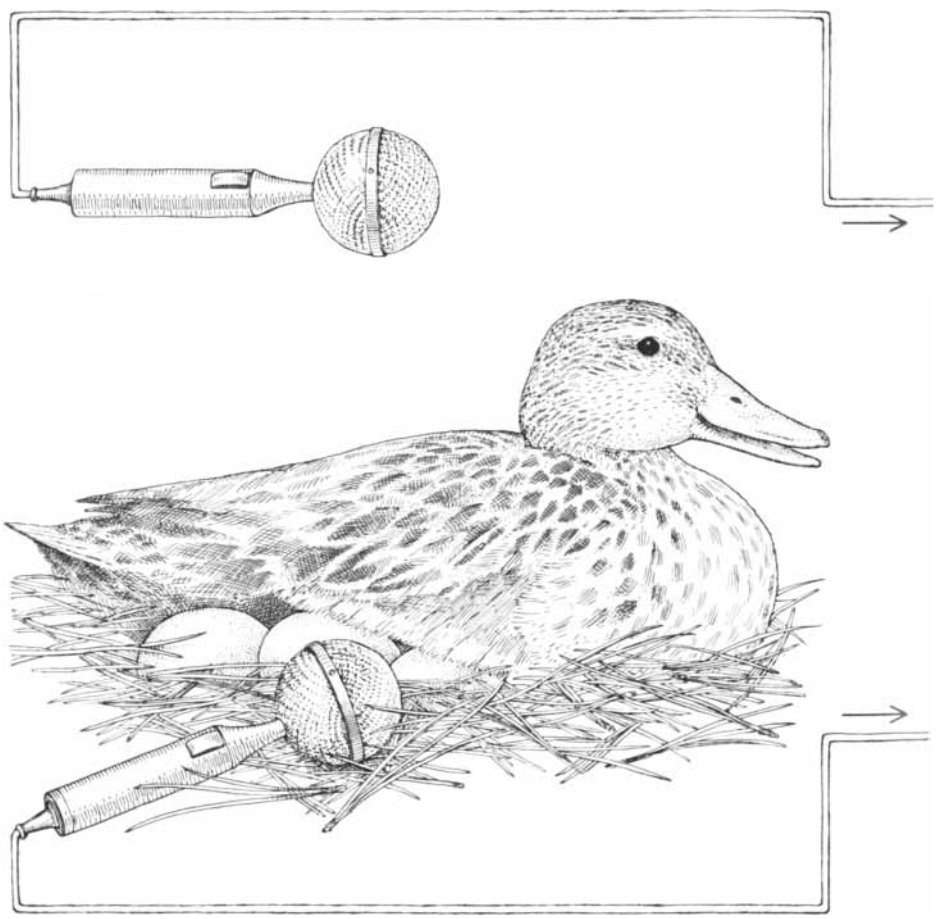
A synthesis of laboratory and field techniques has led to some interesting discoveries about imprinting, the process by which newly hatched birds rapidly form a permanent bond to the parent

by Eckhard H. Hess

In a marsh on the Eastern Shore of Maryland, a few hundred feet from my laboratory building, a female wild mallard sits on a dozen infertile eggs. She has been incubating the eggs for almost four weeks. Periodically she hears the faint peeping sounds that are emitted by hatching mallard eggs, and she clucks softly in response. Since these eggs are infertile, however, they are not about to hatch and they do not emit peeping sounds. The sounds come from a small loudspeaker hidden in the nest under the eggs. The loudspeaker is connected to a microphone next to some hatching mallard eggs inside an incubator in my laboratory. The female mallard can hear any sounds coming from the laboratory eggs, and a microphone beside her relays the sounds she makes to a loudspeaker next to those eggs.

The reason for complicating the life of an expectant duck in such a way is to further our understanding of the phenomenon known as imprinting. It was through the work of the Austrian zoologist Konrad Z. Lorenz that imprinting became widely known. In the 1930's Lorenz observed that newly hatched goslings would follow him rather than their mother if the goslings saw him before they saw her. Since naturally reared geese show a strong attachment for their parent, Lorenz concluded that some animals have the capacity to learn rapidly and permanently at a very early age, and in particular to learn the characteristics of the parent. He called this process of acquiring an attachment to the parent *Prägung*, which in German means "stamping" or "coinage" but in English has been rendered as "imprinting." Lorenz regarded the phenomenon as being different from the usual kind of learning because of its rapidity and apparent permanence. In fact, he was hesitant at first to regard imprinting as a form of learning at all. Some child psychologists and some psychiatrists nevertheless perceived a similarity between the evidence of imprinting in animals and the early behavior of the human infant, and it is not surprising that interest in imprinting spread quickly.

From about the beginning of the 1950's many investigators have intensively studied imprinting in the laboratory. Unlike Lorenz, the majority of them have regarded imprinting as a form of learning and have used methods much the same as those followed in the study of associative learning processes. In every case efforts were made to manipulate or stringently control the imprinting process. Usually the subjects are incuba-



CLUCKS emitted by a female wild mallard in the fourth week of incubating eggs are shown in the sound spectrogram (*upper illustration*). Each cluck lasts for about 150 milliseconds

tor-hatched birds that are reared in the laboratory. The birds are typically kept isolated until the time of the laboratory imprinting experience to prevent interaction of early social experience and the imprinting experience. Various objects have been used as artificial parents: duck decoys, stuffed hens, dolls, milk bottles, toilet floats, boxes, balls, flashing lights and rotating disks. Several investigators have constructed an automatic imprinting apparatus into which the newly hatched bird can be put. In this kind of work the investigator does not observe the young bird directly; all the bird's movements with respect to the imprinting object are recorded automatically.

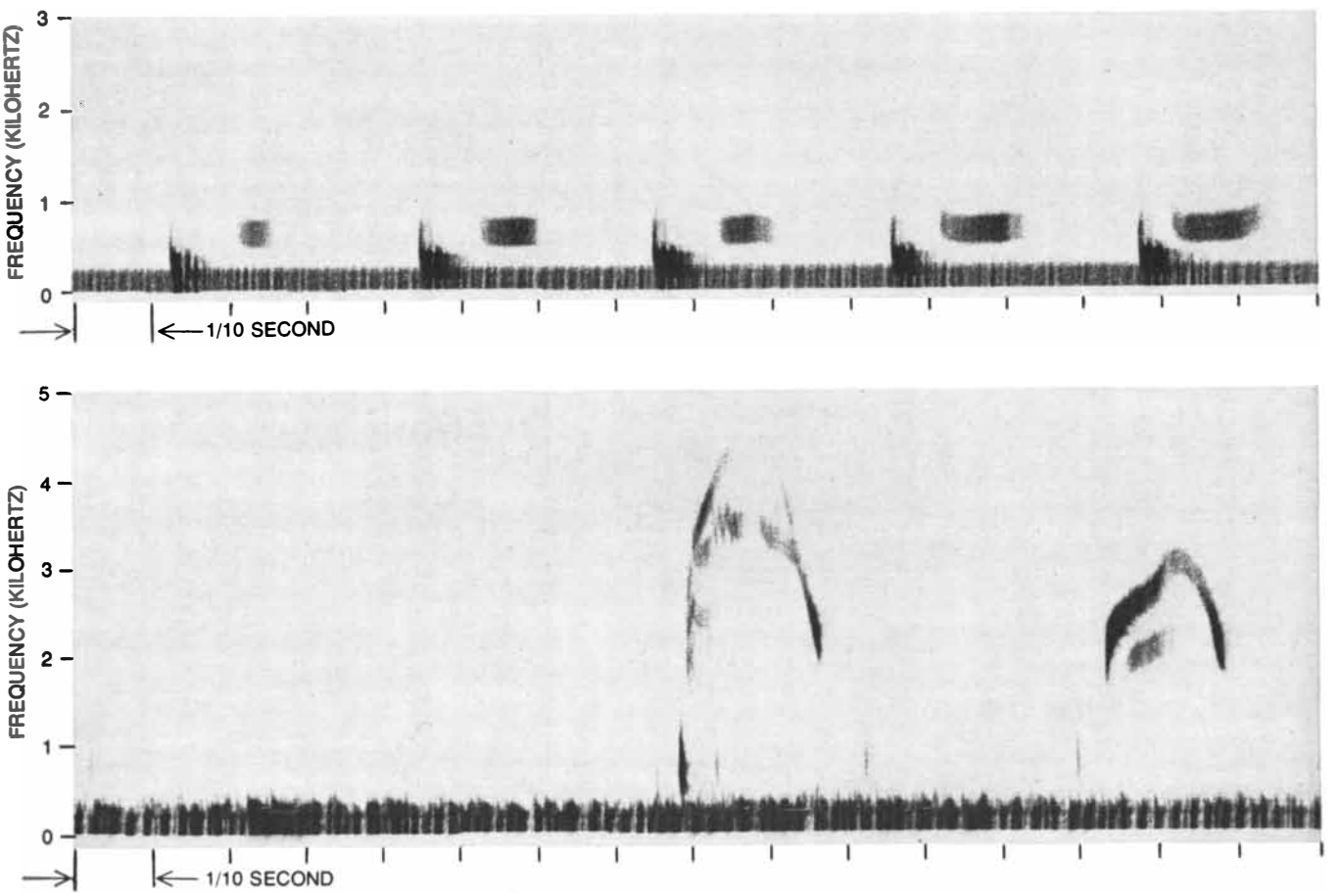
Much of my own research during the past two decades has not differed substantially from this approach. The birds I have used for laboratory imprinting studies have all been incubated, hatched and reared without the normal social and environmental conditions and have then been tested in an artificial situation. It is therefore possible that the behavior observed under such conditions is not relevant to what actually happens in nature.

It is perhaps not surprising that studies of "unnatural" imprinting have produced conflicting results. Lorenz' original statements on the permanence of natural imprinting have been disputed. In many instances laboratory imprinting experiences do not produce permanent and exclusive attachment to the object selected as an artificial parent. For example, a duckling can spend a considerable amount of time following the object to which it is to be imprinted, and immediately after the experience it will follow a completely different object.

In one experiment in our laboratory we attempted to imprint ducklings to ourselves, as Lorenz did. For 20 continuous hours newly hatched ducklings were exposed to us. Before long they followed us whenever we moved about. Then they were given to a female mallard that had hatched a clutch of ducklings several hours before. After only an hour and a half of exposure to the female mallard and other ducklings the human-imprinted ducklings followed the female on the first exodus from the nest. Weeks later the behavior of the human-imprinted ducks was no different from the behavior of the ducks that had been hatched in the nest. Clearly laboratory imprinting is reversible.

We also took wild ducklings from their natural mother 16 hours after hatching and tried to imprint them to humans. On the first day we spent many hours with the ducklings, and during the next two months we made lengthy attempts every day to overcome the ducklings' fear of us. We finally gave up. From the beginning to the end the ducks remained wild and afraid. They were released, and when they had matured, they were observed to be as wary of humans as normal wild ducks are. This result suggests that natural imprinting, unlike artificial laboratory imprinting, is permanent and irreversible. I have had to conclude that the usual laboratory imprinting has only a limited resemblance to natural imprinting.

It seems obvious that if the effects of natural imprinting are to be understood, the phenomenon must be studied as it

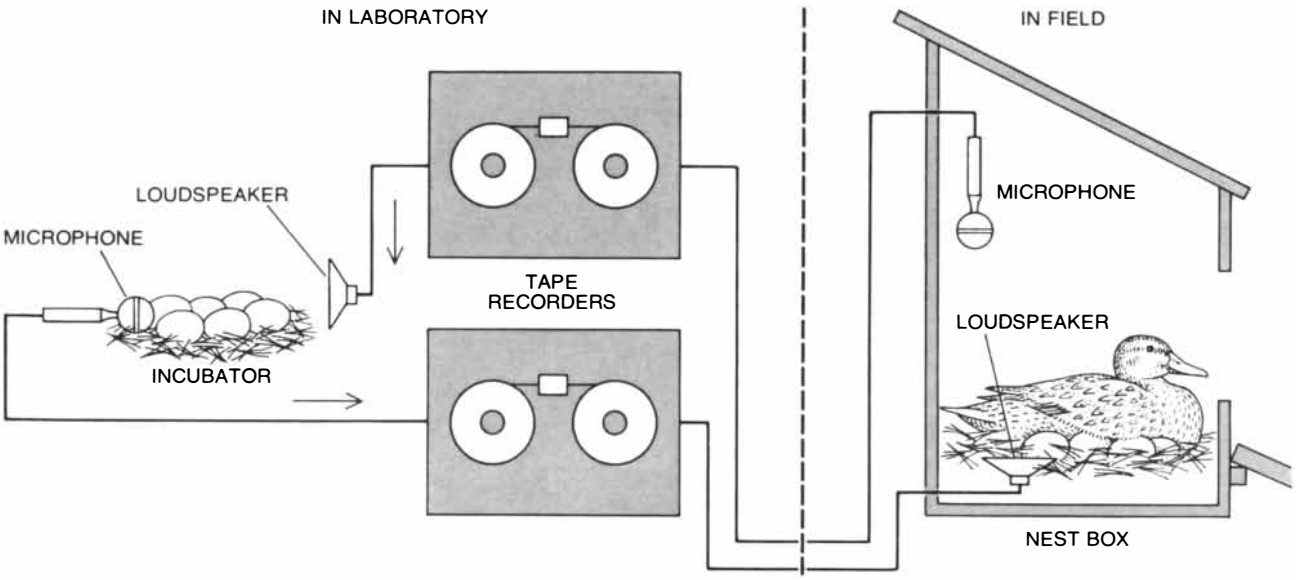


and is low in pitch: about one kilohertz or less. Sounds emitted by ducklings inside the eggs are high-pitched, rising to about four kilohertz (*lower illustration*). Records of natural, undisturbed imprinting events in the nest provide a control for later experiments.

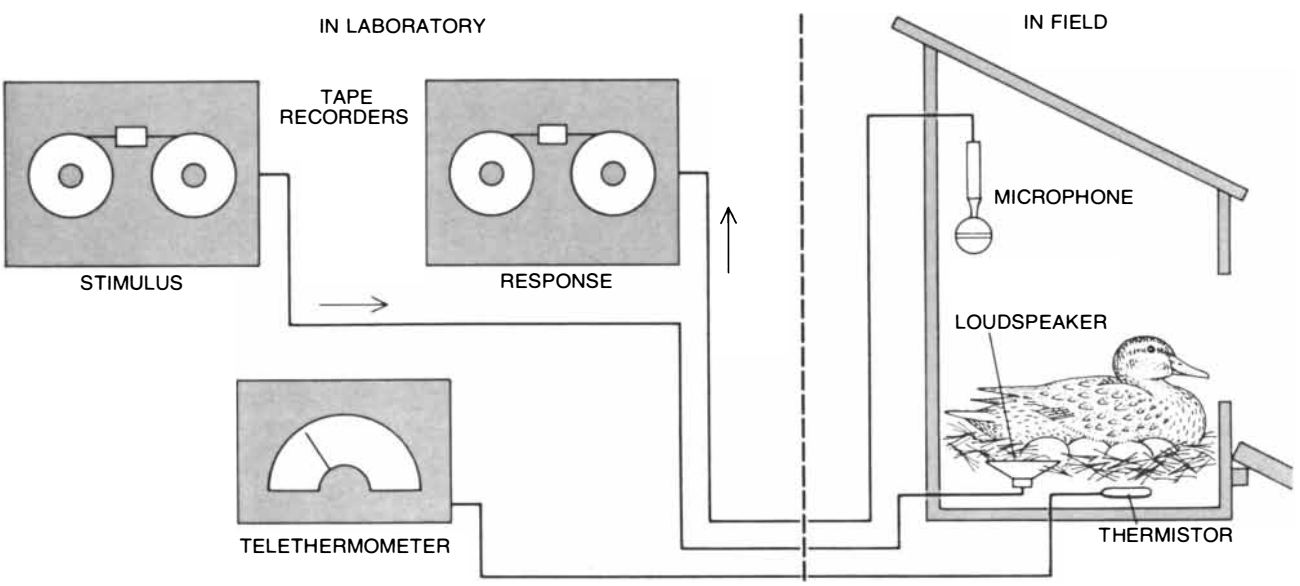
operates in nature. The value of such studies was stressed as long ago as 1914 by the pioneer American psychologist John B. Watson. He emphasized that field observations must always be made to test whether or not conclusions drawn from laboratory studies conform to what actually happens in nature. The disparity between laboratory results and what happens in nature often arises from the failure of the investigator to really look at the animal's behavior. For years I have cautioned my students against shutting their experimental animals in "black boxes" with automatic recording devices and never directly observing how the animals behave.

This does not mean that objective laboratory methods for studying the behavior of animals must be abandoned. With laboratory investigations large strides have been made in the development of instruments for the recording of behavior. In the study of imprinting it is not necessary to revert to imprecise naturalistic observations in the field. We can now go far beyond the limitations of traditional field studies. It is possible to set up modern laboratory equipment in actual field conditions and in ways that do not disturb or interact with the behavior being studied, in other words, to achieve a synthesis of laboratory and field techniques.

The first step in the field-laboratory method is to observe and record the undisturbed natural behavior of the animal in the situation being studied. In our work on imprinting we photographed the behavior of the female mallard during incubation and hatching. We photographed the behavior of the ducklings



FEMALE MALLARD sitting on infertile eggs hears sounds transmitted from mallard eggs in a laboratory incubator. Any sounds she makes are transmitted to a loudspeaker beside the eggs in the laboratory. Such a combination of field and laboratory techniques permits recording of events without disturbing the nesting mallard and provides the hatching eggs with nearly natural conditions.



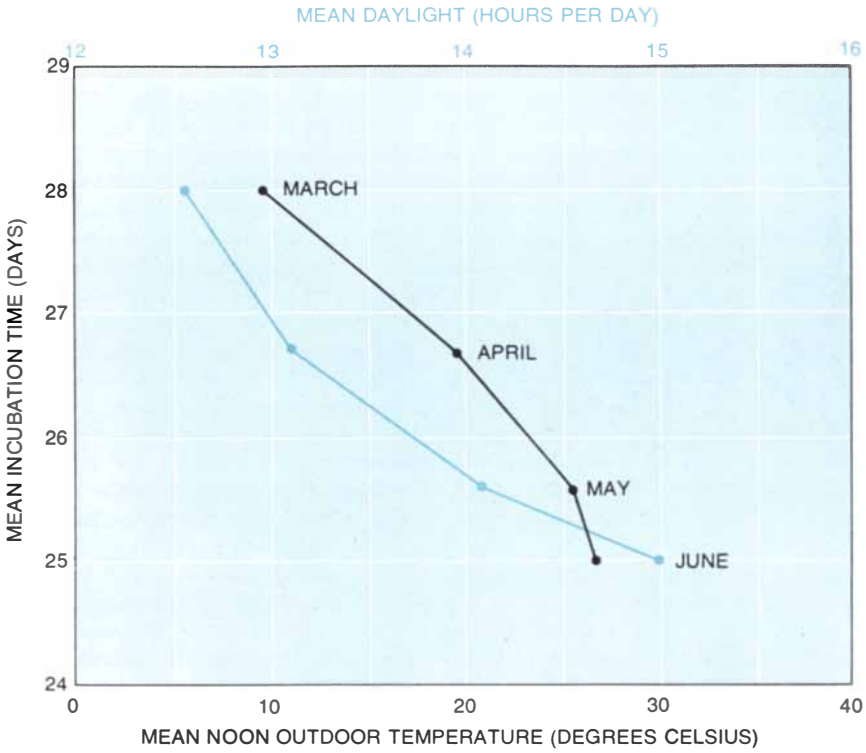
REMOTE MANIPULATION of prehatching sounds is accomplished by placing a sensitive microphone and a loudspeaker in the nest of a female wild mallard who is sitting on her own eggs. Prerecorded hatching-duckling sounds are played at specified times through the loudspeaker and the female mallard's responses to this stimulus are recorded. A thermistor probe transmits the temperature in the nest to a telethermometer and chart recorder. The thermistor records provide data about when females are on nest.

during and after hatching. We recorded all sounds from the nest before and after hatching. Other factors, such as air temperature and nest temperature, were also recorded.

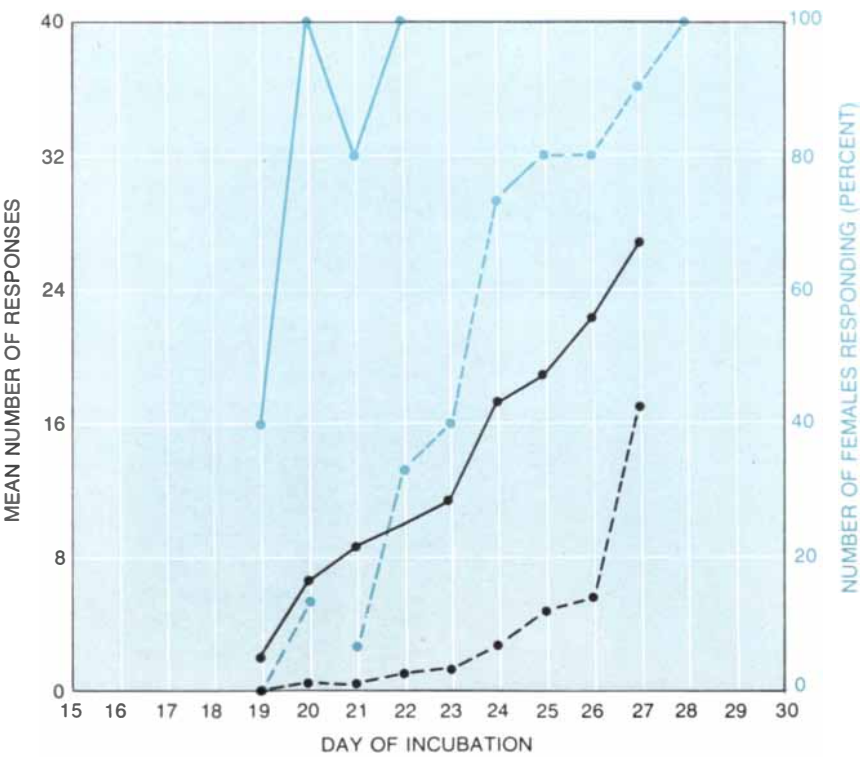
A detailed inventory of the actual events in natural imprinting is essential for providing a reference point in the assessment of experimental manipulations of the imprinting process. That is, the undisturbed natural imprinting events form the control situation for assessing the effects of the experimental manipulations. This is quite different from the "controlled" laboratory setting, in which the ducklings are reared in isolation and then tested in unnatural conditions. The controlled laboratory study not only introduces new variables (environmental and social deprivation) into the imprinting situation but also it can prevent the investigator from observing factors that are relevant in wild conditions.

My Maryland research station is well suited for the study of natural imprinting in ducks. The station, near a national game refuge, has 250 acres of marsh and forest on a peninsula on which there are many wild and semiwild mallards. Through the sharp eyes of my technical assistant Elihu Abbott, a native of the Eastern Shore, I have learned to see much I might otherwise have missed. Initially we looked at and listened to the undisturbed parent-offspring interaction of female mallards that hatched their own eggs both in nests on the ground and in specially constructed nest boxes. From our records we noticed that the incubation time required for different clutches of eggs decreased progressively between March and June. Both the average air temperature and the number of daylight hours increase during those months; both are correlated with the incubation time of mallard eggs. It is likely, however, that temperature rather than photoperiod directly influences the duration of incubation. In one experiment mallard eggs from an incubator were slowly cooled for two hours a day in a room with a temperature of seven degrees Celsius, and another set of eggs was cooled in a room at 27 degrees C. These temperatures respectively correspond to the mean noon temperatures at the research station in March and in June. The eggs that were placed in the cooler room took longer to hatch, indicating that temperature affects the incubation time directly. Factors such as humidity and barometric pressure may also play a role.

We noticed that all the eggs in a wild



INCUBATION TIME of mallard eggs hatched naturally in a feral setting at Lake Cove, Md., decreased steadily from March to June. The incubation period correlated with both the outdoor temperature (*black curve*) and the daily photoperiod (*colored curve*).



VOCAL RESPONSES to hatching-duckling sounds of 15 female wild mallards (*broken curves*) and five human-imprinted mallards (*solid curves*), which were later released to the wild, followed the same pattern, although the human-imprinted mallards began responding sooner and more frequently. A tape recording of the sounds of a hatching duckling was played daily throughout the incubation period to each female mallard while she was on her nest. Responses began on the 19th day of incubation and rose steadily until hatching.

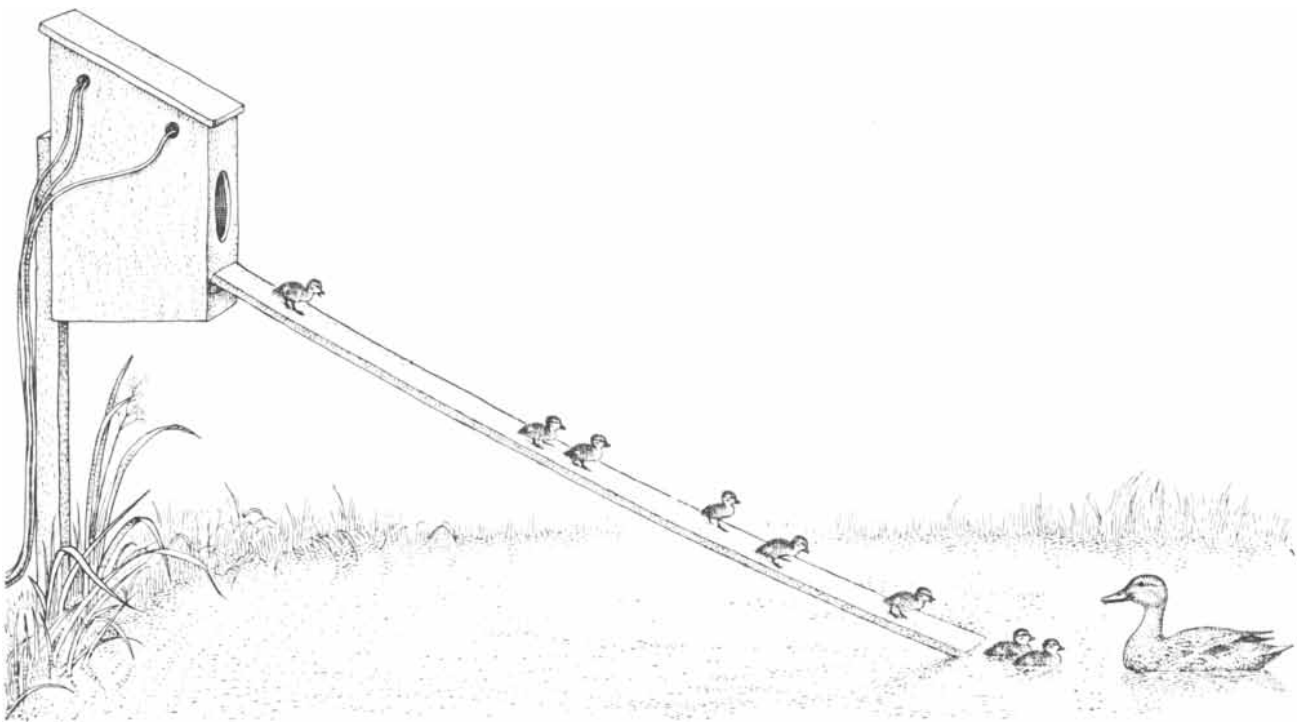
nest usually hatch between three and eight hours of one another. As a result all the ducklings in the same clutch are approximately the same age in terms of the number of hours since hatching. Yet when mallard eggs are placed in a mechanical incubator, they will hatch over a two- or three-day period even when precautions are taken to ensure that all the eggs begin developing simultaneously. The synchronous hatching observed in nature obviously has some survival value. At the time of the exodus from the nest, which usually takes place between 16 and 32 hours after hatching, all the ducklings would be of a similar age and thus would have equal motor capabilities and similar social experiences.

Over the years our laboratory studies and actual observations of how a female mallard interacts with her offspring have pointed to the conclusion that imprinting is related to the age after hatching rather than the age from the beginning of incubation. Many other workers, however, have accepted the claim that age from the beginning of incubation determines the critical period for maximum effectiveness of imprinting. They base their belief on the findings of Gilbert Gottlieb of the Dorothea Dix Hospital in Raleigh, N.C., who in a 1961 paper described experiments that apparently showed that maximum imprinting in ducklings occurs in the period between 27 and 27½ days after the beginning of incubation. To make sure that all the eggs he was working with started incubation at the same time he first chilled the eggs so that any partially developed embryos would be killed. Yet the 27th day after the beginning of incubation can hardly be the period of maximum imprinting for wild ducklings that hatch in March under natural conditions, because such ducklings take on the average 28 days to hatch. Moreover, if the age of a duckling is measured from the beginning of incubation, it is hard to explain why eggs laid at different times in a hot month in the same nest will hatch within six to eight hours of one another under natural conditions.

Periodic cooling of the eggs seems to affect the synchronization of hatching. The mallard eggs from an incubator that were placed in a room at seven degrees C. hatched over a period of a day and a half, whereas eggs placed in the room at 27 degrees hatched over a period of two and a half days (which is about normal for artificially incubated eggs). Cooling cannot, however, play a major role. In June the temperature in the outdoor nest boxes averages close to the normal brooding temperature while the female mallard is absent. Therefore an egg laid on June 1 has a head start in incubation over those laid a week later. Yet we have observed that all the eggs in clutches laid in June hatch in a period lasting between six and eight hours.

We found another clue to how the synchronization of hatching may be achieved in the vocalization pattern of the brooding female mallard. As many others have noted, the female mallard vocalizes regularly as she sits on her eggs during the latter part of the incubation period. It seemed possible that she was vocalizing to the eggs, perhaps in response to sounds from the eggs themselves. Other workers had observed that ducklings make sounds before they hatch, and the prehatching behavior of ducklings in response to maternal calls has been extensively reported by Gottlieb.

We placed a highly sensitive microphone next to some mallard eggs that were nearly ready to hatch. We found that the ducklings indeed make sounds while they are still inside the egg. We made a one-minute tape recording of the sounds emitted by a duckling that had pipped its shell and was going to hatch within the next few hours. Then we made a seven-minute recording that would enable us to play the duckling sounds three times for one minute interspersed with one-minute silences. We played the recording once each to 37 female mallards at various stages of incubation. There were no positive responses from the female mallards during



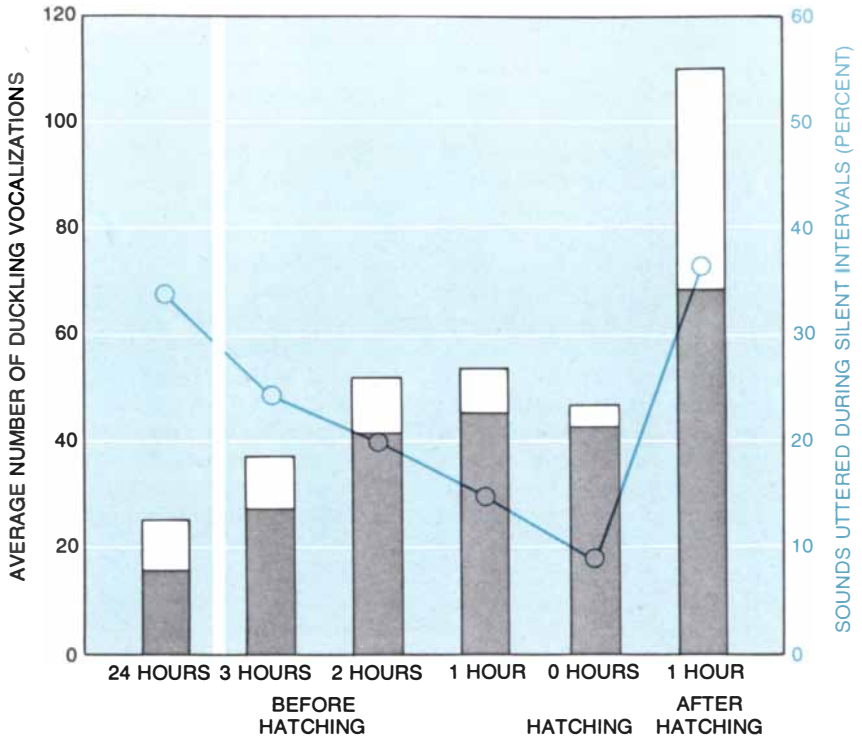
NEST EXODUS takes place about 16 to 32 hours after hatching. The female mallard begins to make about 40 to 65 calls per minute and continues while the ducklings leave the nest to follow her. The ducklings are capable of walking and swimming from hatching.

the first and second week of incubation. In fact, during the first days of incubation some female mallards responded with threat behavior: a fluffing of the feathers and a panting sound. In the third week some females responded to the recorded duckling sounds with a few clucks. In the fourth week maternal clucks were frequent and were observed in all ducks tested.

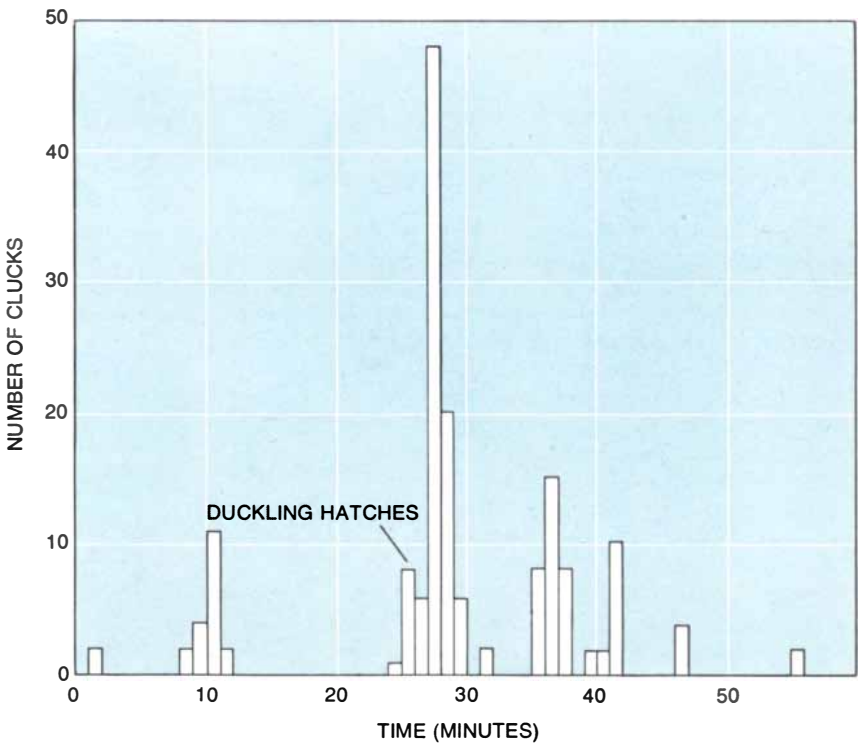
We found the same general pattern of response whether the female mallards were tested once or, as in a subsequent experiment, tested daily during incubation. Mallards sitting on infertile eggs responded just as much to the recorded duckling sounds as mallards sitting on fertile eggs did. Apparently after sitting on a clutch of eggs for two or three weeks a female mallard becomes ready to respond to the sounds of a hatching duckling. There is some evidence that the parental behavior of the female mallard is primed by certain neuroendocrine mechanisms. We have begun a study of the neuroendocrine changes that might accompany imprinting and filial behavior in mallards.

To what extent do unhatched ducklings respond to the vocalization of the female mallard? In order to find out we played a recording of a female mallard's vocalizations to ducklings in eggs that had just been pipped and were scheduled to hatch within the next 24 hours. As before, the sounds were interspersed with periods of silence. We then recorded all the sounds made by the ducklings during the recorded female mallard vocalizations and also during the silent periods on the tape. Twenty-four hours before the scheduled hatching the ducklings emitted 34 percent of their sounds during the silent periods, which suggests that at this stage they initiate most of the auditory interaction. As hatching time approaches the ducklings emit fewer and fewer sounds during the silent periods. The total number of sounds they make, however, increases steadily. At the time of hatching only 9 percent of the sounds they make are emitted during the silent periods. One hour after hatching, in response to the same type of recording, the ducklings gave 37 percent of their vocalizations during the silent periods, a level similar to the level at 24 hours before hatching.

During the hatching period, which lasts about an hour, the female mallard generally vocalizes at the rate of from zero to four calls per one-minute interval. Occasionally there is an interval in which she emits as many as 10 calls. When the duckling actually hatches, the female mallard's vocalization increases



NUMBER OF SOUNDS from ducklings before and after hatching are shown. The ducklings heard a recording consisting of five one-minute segments of a female mallard's clucking sounds interspersed with five one-minute segments of silence. The recording was played to six mallard eggs and the number of vocal responses by the ducklings to the clucking segments (*gray bars*) and to the silent segments (*white bars*) were counted. Twenty-four hours before hatching 34 percent of the duckling sounds were made during the silent interval, indicating the ducklings initiated a substantial portion of the early auditory interaction. As hatching time approached the ducklings initiated fewer and fewer of the sounds and at hatching vocalized most in response to the clucks of the female mallard.



CLUCKING RATE of a wild, ground-nesting female mallard rose dramatically for about two minutes while a duckling hatched and then slowly declined to the prehatching rate. Each bar depicts the number of clucks emitted by the female during a one-minute period.

dramatically to between 45 and 68 calls per minute for one or two minutes.

Thus the sounds made by the female mallard and by her offspring are complementary. The female mallard vocalizes most when a duckling has just hatched. A hatching duckling emits its cries primarily when the female is vocalizing.

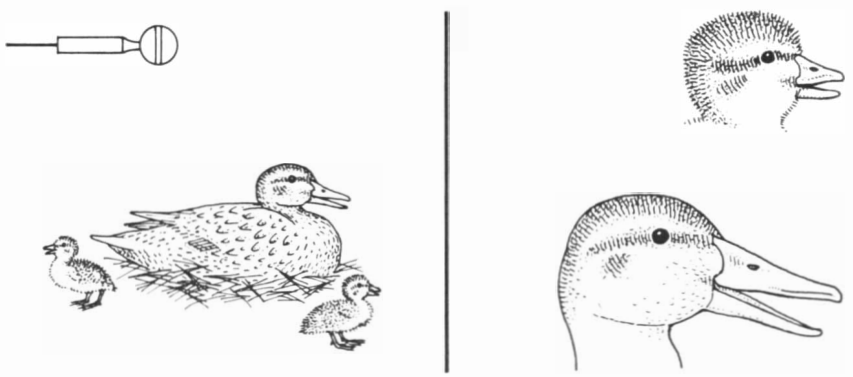
After all the ducklings have hatched the female mallard tends to be relatively quiet for long intervals, giving between zero and four calls per minute. This continues for 16 to 32 hours until it is time for the exodus from the nest. As the exodus begins the female mallard quickly builds up to a crescendo of between 40 and 65 calls per minute; on rare occasions we have observed between 70 and 95 calls per minute. The duration of the high-calling-rate period depends on how quickly the ducklings leave the nest to follow her. There is now a change in the sounds made by the female mallard. Up to this point she has been making clucking sounds. By the time the exodus from the nest takes place some of her sounds are more like quacks.

The auditory interaction of the female mallard and the duckling can begin well before the hatching period. As I have indicated, the female mallard responds to unhatched-duckling sounds during the third and fourth week of incubation. Normally ducklings penetrate a membrane to reach an air space inside the eggshell two days before hatching. We have not found any female mallard that vocalized to her clutch before the duckling in the egg reached the air space. We have found that as soon as the duckling penetrates the air space the female begins to cluck at a rate of between zero and four times per minute. Typically she continues to vocalize at this rate until the ducklings begin to pip their eggs (which is about 24 hours after they have entered the air space). As the eggs are being pipped the female clucks at the rate of between 10 and 15 times per minute. When the pipping is completed, she drops back to between zero and four calls per minute. In the next 24 hours there is a great deal of auditory interaction between the female and her unhatched offspring; this intense interaction may facilitate the rapid formation of the filial bond after hatching, although it is quite possible that synchrony of hatching is the main effect. Already we have found that a combination of cooling the eggs daily, placing them together so that they touch one another and transmitting parent-young vocal responses through the microphone-loudspeaker hookup between the female's nest and the laboratory incubator causes the eggs in the incubator to hatch as synchronously as eggs in nature do. In fact, the two times we did this we found that all the eggs in the clutches hatched within four hours of one another. It has been shown in many studies of imprinting, including laboratory studies, that auditory stimuli have an important effect on the development of filial attachment. Auditory stimulation, before and after hatching, together with tactile stimulation in the nest after hatching results in ducklings that are thoroughly imprinted to the female mallard that is present.

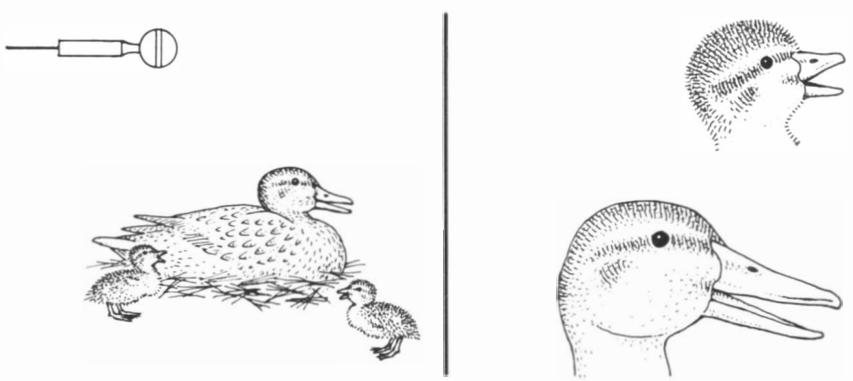
Furthermore, it appears that auditory interaction before hatching may play an important role in promoting the synchronization of hatching. As our experiments showed, not only does the female mallard respond to sounds from her eggs but also the ducklings respond to her clucks. Perhaps the daily cooling of the eggs when the female mallard leaves the nest to feed serves to broadly synchronize embryonic and behavioral development, whereas the auditory interaction of the mother with the ducklings and of one duckling with another serves to provide finer synchronization. Margaret Vince of the University of Cambridge has shown that the synchronization of hatching in quail is promoted by the mutual auditory interaction of the young birds in the eggs.

Listening to the female mallards vocalize to their eggs or to their newly hatched offspring, we were struck by the fact that we could tell which mallard was vocalizing, even when we could not see her. Some female mallards regularly emit single clucks at one-second intervals, some cluck in triple or quadruple clusters and others cluck in clusters of different lengths. The individual differences in the vocalization styles of female mallards may enable young ducklings to identify their mother. We can also speculate that the characteristics of a female mallard's voice are learned by her female offspring, which may then adopt a similar style when they are hatching eggs of their own.

The female mallards not only differ from one another in vocalization styles but also emit different calls in different situations. We have recorded variations in pitch and duration from the same mallard in various nesting situations. It



SOUND SPECTROGRAM of the calls of newly hatched ducklings in the nest and the mother's responses is shown at right. The high-pitched peeps of the ducklings are in the



DISTRESS CALLS of ducklings in the nest evoke a quacklike response from the female mallard. The cessation of the distress calls and the onset of normal duckling peeping sounds

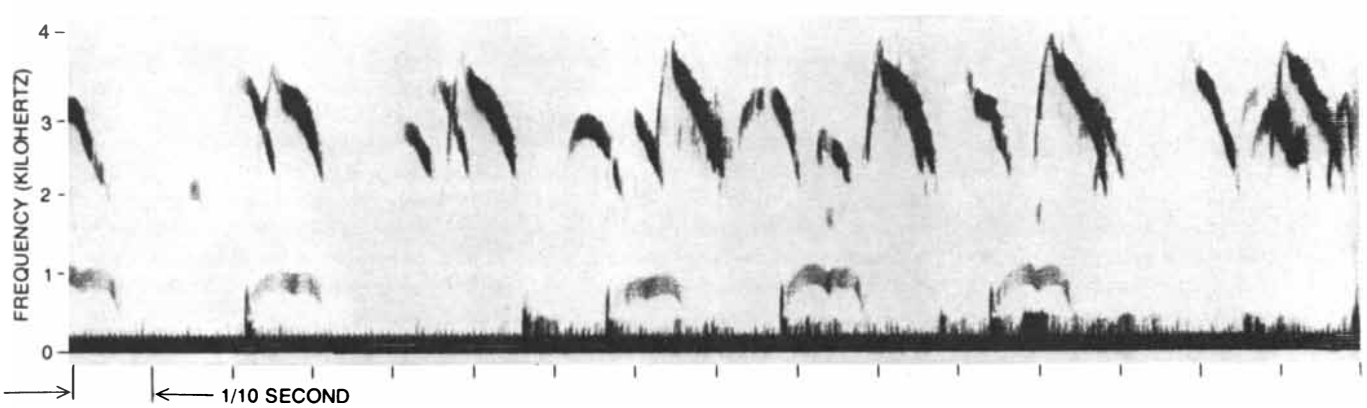
seems likely that such variations in the female mallard call are an important factor in the imprinting process.

Studies of imprinting in the laboratory have shown that the more effort a duckling has to expend in following the imprinting object, the more strongly it prefers that object in later testing. At first it would seem that this is not the case in natural imprinting; young ducklings raised by their mother have little difficulty following her during the exodus from the nest. Closer observation of many nests over several seasons showed, however, that ducklings make a considerable effort to be near their parent. They may suffer for such efforts, since they can be accidentally stepped on, squeezed or scratched by the female adult. The combination of effort and punishment may actually strengthen imprinting. Work in my laboratory showed that chicks given an electric shock while they were following the imprinting object later showed stronger attachment to the object than unshocked chicks did. It is reasonable to expect similar results with ducklings.

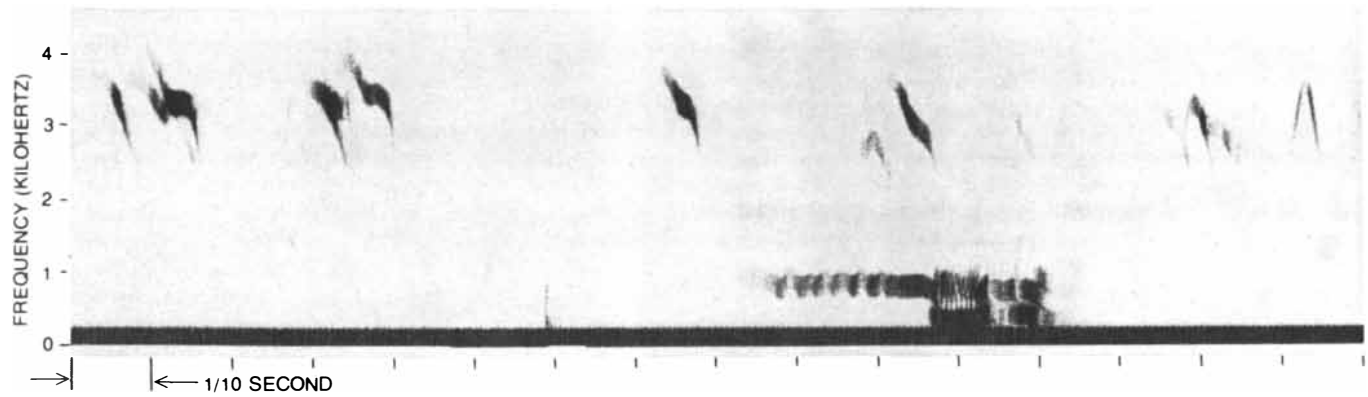
Slobodan Petrovich of the University of Maryland (Baltimore County) and I have begun a study to determine the relative contributions of prehatching and posthatching auditory experience on imprinting and filial attachment. The auditory stimuli consist of either natural mallard maternal clucks or a human voice saying "Come, come, come." Our results indicate that prehatching stimulation by natural maternal clucks may to a degree facilitate the later recognition of the characteristic call of the mallard. Ducklings lacking any experience with a maternal call imprint as well to a duck decoy that utters "Come, come, come" as to a decoy that emits normal mallard clucks. Ducklings that had been exposed to a maternal call before hatching imprinted better to decoys that emitted the mallard clucks. We found, however, that the immediate posthatching experiences, in this case with a female mallard on the nest, can highly determine the degree of filial attachment and make imprinting to a human sound virtually impossible.

It is important to recognize that almost all laboratory imprinting experiments, including my own, have been deprivation experiments. The justification for such experiments has been the ostensible need for controlling the variables of the phenomenon, but the deprivation may have interfered with the normal behavioral development of the young ducklings. Whatever imprinting experiences the experimenter allows therefore do not produce the maximum effect.

Although our findings are far from complete, we have already determined enough to demonstrate the great value of studying imprinting under natural conditions. The natural laboratory can be profitably used to study questions about imprinting that have been raised but not answered by traditional laboratory experiments. We must move away from the in vitro, or test-tube, approach to the study of behavior and move toward the in vivo method that allows interaction with normal environmental factors. Some of the questions are: What is the optimal age for imprinting? How long must the imprinting experience last for it to have the maximum effect? Which has the greater effect on behavior: first experience or the most recent experience? Whatever kind of behavior is being studied, the most fruitful approach may well be to study the behavior in its natural context.



two-to-four-kilohertz range. They normally have the shape of an inverted *V*. The female mallard's clucks are about one kilohertz and last about 130 milliseconds. After the eggs hatch the vocalization of the female changes both in quantity and in quality of sound.



is almost immediate, as can be seen in this sound spectrogram. The female mallard's quacklike call is about one kilohertz in pitch and has a duration of approximately 450 milliseconds. The call is emitted about once every two seconds in response to distress cries.

THE NATURE OF AROMATIC MOLECULES

"Aromaticity" refers to the exceptional stability of certain ring-shaped organic molecules. Whether a given system is aromatic, nonaromatic or antiaromatic depends on how many of its electrons are "delocalized"

by Ronald Breslow

In common with the other natural sciences chemistry is fundamentally concerned with understanding the laws of nature. Part of the effort is directed at understanding general laws that govern the common behavior of all matter at the molecular level. Most basic chemical research, however, is directed at understanding the properties of individual chemical compounds and how molecular structure determines these properties. In trying to correlate structure with properties the chemist is not limited to substances presented to him by nature. Perhaps more than investigators in any other discipline, the chemist is concerned with the properties of novel substances and substances prepared artificially. Chemistry remains a "natural" science, but it is concerned not only with structures and phenomena preexisting in nature but also with innumerable other structures and phenomena that are possible through the operation of nature's chemical laws.

For this reason synthesis has always been an important activity in chemical investigation. The modern chemical industry abundantly illustrates the practical results that can flow from the synthesis of new chemical substances with unusual properties. The large majority of these substances owe their existence to the work of investigators who were interested only in exploring the natural laws of chemistry and who perceived that the synthesis and study of new structures were more rewarding than simple examination of those nature has already provided. The rewards of this approach are well illustrated by the history of the development of the concepts of aromaticity and antiaromaticity in organic chemistry.

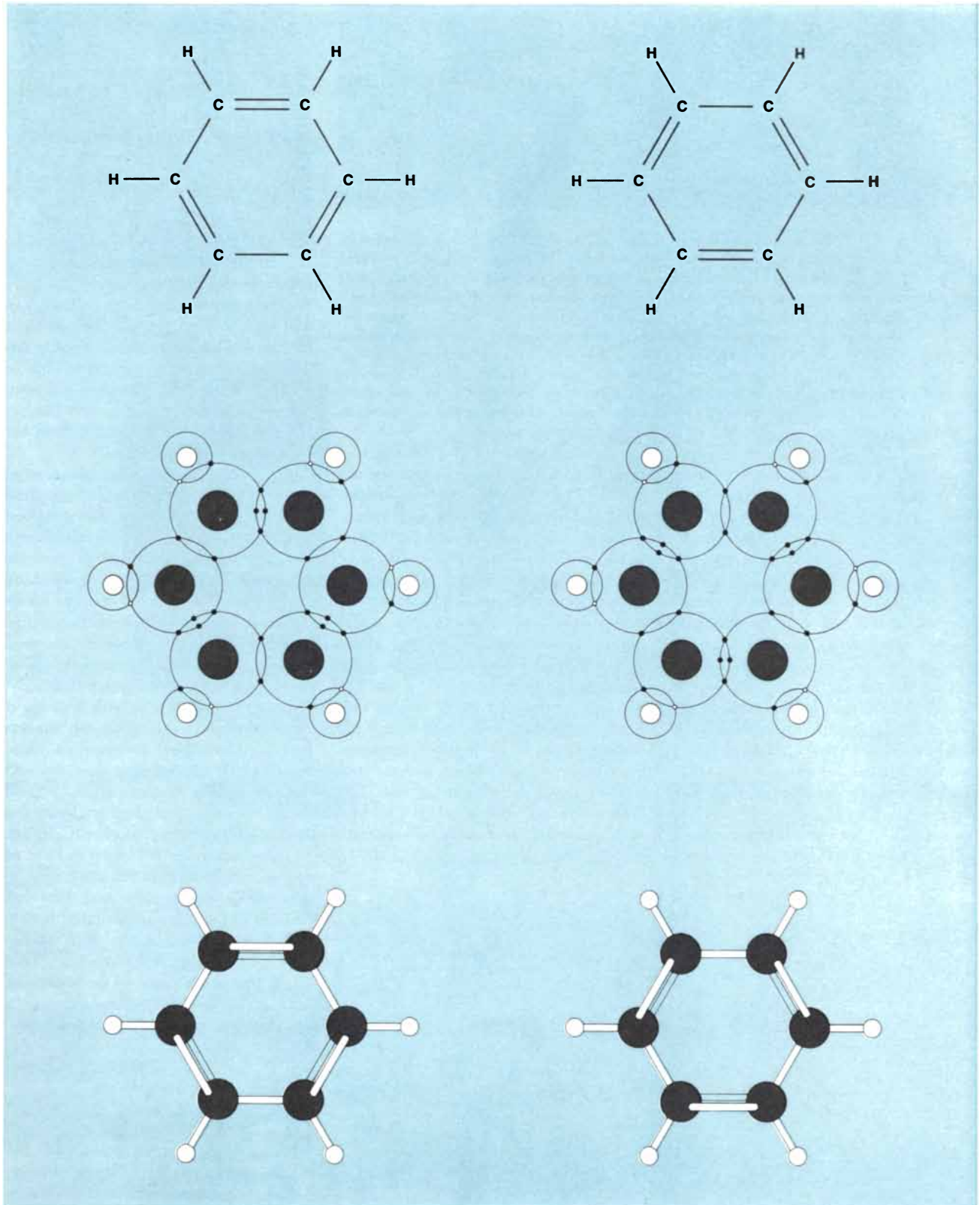
Chemistry is cursed with unsuitable ancient words and names that are understood by practitioners and confuse the uninitiated. "Organic" chemistry is the name for any chemistry concerned with carbon compounds. Before the 1820's it was believed the carbon compounds that can be isolated from living organisms could only be made "organically." No one has doubted for a long time that chemists can synthesize all such compounds in the laboratory, but the name lingers on.

By the same token it was recognized early that the compound benzene and substances related to it had special properties. One of the most obvious was that such compounds had an odor different from that of other carbon compounds; hence the group of substances including benzene was called "aromatic." The aroma of these substances is the least interesting special thing about them, but unfortunately chemists then used the term "aromaticity" to describe the much more interesting chemical properties of benzene and its relatives. The term is so deeply embedded in the chemical literature that it would be fruitless to try to change it now, but clearly what the chemist understands by aromaticity has nothing to do with the layman's interpretation of the word.

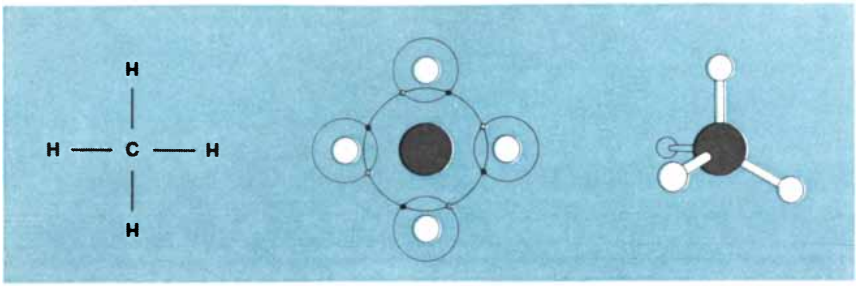
Benzene is a colorless liquid composed of molecules with the formula C_6H_6. The six carbon atoms of benzene are arranged in a regular hexagon; each carbon atom has one hydrogen atom attached to it by a single covalent bond consisting of two electrons, one supplied by the carbon atom and the other by the hydrogen atom [*see illustration on opposite page*]. In order to appreciate why the chemical properties of benzene were surprising, one must consider what is known about other hydrocarbons (compounds consisting only of carbon and hydrogen). The simplest hydrocarbon is methane (CH_4) [*see top illustration on page 34*]. In this molecule the four hydrogen atoms are attached to the central carbon atom by four covalent bonds, each consisting of two shared electrons between the carbon atom and each hydrogen atom. Therefore each hydrogen atom has a filled electronic valence shell containing two electrons: the two it is sharing with a carbon atom. The carbon atom also has a filled valence shell of eight electrons, consisting of four of its own electrons together with the four provided by the four hydrogen atoms. (The carbon atom also has two electrons bound close to its nucleus in an inner shell, but these electrons do not participate in chemical reactions.)

Stable compounds of carbon almost always have eight electrons surrounding carbon in its valence shell; when the eight electrons participate in four single bonds to another atom, so that each single bond incorporates two shared electrons, one says that the carbon is saturated. The molecule ethane (C_2H_6) is also a saturated hydrocarbon, since it has eight electrons around each carbon in the form of four two-electron single bonds. In ethane, however, one of the single bonds joins the two carbon atoms.

In contrast, the molecule ethylene (C_2H_4) is the simplest member of the class of unsaturated hydrocarbons [*see bottom illustration on page 34*]. In this molecule each carbon atom is attached to two hydrogen atoms by two-electron single bonds, but the two central carbon atoms are held together by a double bond, which means that two electron pairs are shared by these carbons. The result again is that each hydrogen shares one electron pair and has a filled valence shell, whereas each carbon shares four electron pairs and also has a filled valence shell. The difference is simply that in ethylene two of the electron pairs are shared by the same two atoms.



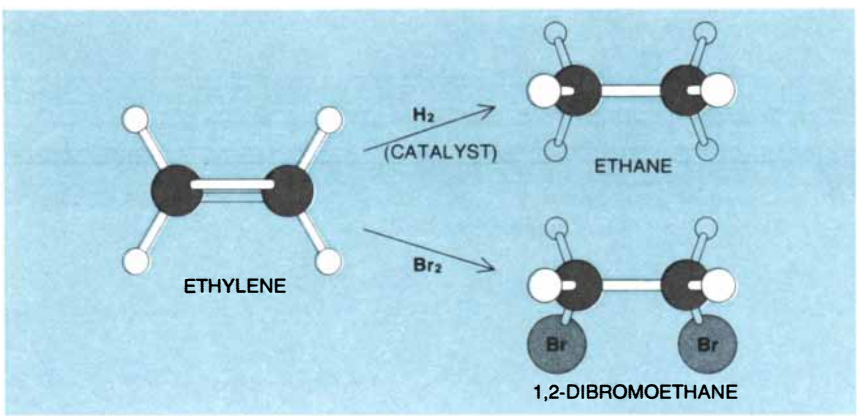
BENZENE, a colorless liquid composed of molecules with the formula C_6H_6, is the historical prototype of the aromatic compounds. The six carbon atoms of benzene all lie in the same plane, situated at the corners of a regular hexagon. Each carbon atom has one hydrogen atom attached to it by a single covalent bond consisting of two electrons, one supplied by the carbon atom and the other by the hydrogen atom. The carbon atoms are joined to form the closed benzene ring by a combination of three single bonds and three double bonds. The positions of the double bonds can be assigned arbitrarily in two different ways (*left*, *right*), and the true structure is regarded as a hybrid of these two resonance forms. Each form is in turn represented in this illustration in three different ways, beginning with the familiar chemical diagrams (*top*). In the electron-shell diagrams (*middle*) electrons contributed by the hydrogen atoms are indicated by white dots; electrons contributed by carbon atoms are indicated by black dots. (Only the six valence, or outer-shell, electrons of each carbon are shown.) Ball-and-stick convention (*bottom*) is used in most of the following illustrations.



METHANE (CH_4), the simplest hydrocarbon, is composed of four hydrogen atoms attached to the central carbon atom by four covalent bonds, each consisting of two shared electrons. As the electron-shell diagram shows, each hydrogen atom has a filled valence shell containing two electrons (its one electron together with one provided by the carbon atom), whereas each carbon atom has a filled valence shell of eight electrons (four of its own electrons together with the four provided by the four hydrogen atoms). Stable hydrocarbon compounds of this type, in which the eight valence electrons of carbon participate in four single bonds to another atom, so that each single bond incorporates two shared electrons, are said to be saturated. As the ball-and-stick diagram shows, the four hydrogens of the methane molecule lie at the corners of a regular tetrahedron, with the carbon at the center.

Long experience has made it clear that such double bonds are more easily broken than single bonds. Hence it is fairly easy to react unsaturated molecules with a variety of chemical reagents. For example, in the presence of a catalyst such as platinum or nickel one molecule of ethylene will readily react with one molecule of hydrogen (H_2), a process that breaks and eliminates one of the two carbon-carbon bonds. Ethylene is then chemically transformed into ethane. This process, catalytic hydrogenation, is widely used to convert relatively reactive unsaturated compounds into saturated ones by the addition of hydrogen atoms at the site of carbon-carbon double bonds. In the manufacture of margarine or of peanut butter it is common practice to hydrogenate some of the unsaturated fats, thereby removing at least some of the carbon-carbon double bonds.

Unsaturated compounds participate in other chemical reactions with great ease. For instance, ethane is essentially inert when it is mixed with bromine molecules (Br_2), whereas ethylene reacts rapidly to add the two bromine atoms across the carbon-carbon double bond and again produce a compound containing only saturated carbon, in this case 1,2-dibromoethane. (By numbering the carbon atoms in an organic compound the chemist can tell where various substituent atoms are attached; thus "1,2-dibromoethane" tells him that two bromine atoms are attached to two separate carbon atoms.)



ETHYLENE (C_2H_4), the simplest member of the class of unsaturated hydrocarbons, contains two central carbon atoms held together by an easily broken double bond. Such unsaturated molecules react quite readily with a variety of chemical reagents. Two typical reactions are shown here. In the process called catalytic hydrogenation one molecule of ethylene (*left*) reacts with one molecule of hydrogen (H_2) in the presence of a catalyst; the reaction eliminates one of the two carbon-carbon bonds, chemically transforming the ethylene in ethane (*top right*). In the presence of bromine molecules (Br_2) ethylene reacts rapidly to add the two bromine atoms across the carbon-carbon double bond, again producing a compound containing only saturated carbon, in this case 1,2-dibromoethane (*bottom right*).

Returning now to the benzene molecule, we can first of all describe the kind of structure we expect it to have and then see how its properties differ from those one might expect by simple analogy with more familiar compounds. We know that each carbon atom of the benzene molecule can form only a single bond to a neighboring hydrogen atom. If each carbon formed only single bonds to each of its two neighboring carbons, it would have only three single bonds, accounting for only six electrons. We conclude that benzene needs one more bond at each of its carbons, so that each will have a filled valence shell of eight electrons. This requirement is easily satisfied if we simply put three double bonds into the molecule: a double bond between carbons 1 and 2, another between carbons 3 and 4 and a third between carbons 5 and 6.

Early chemists drew this unsaturated structure but were troubled because benzene did not behave like a typical unsaturated molecule [*see illustration on opposite page*]. Although benzene can be hydrogenated to cyclohexane (C_6H_{12}) by the addition of three molecules of hydrogen, the reaction requires significantly higher temperatures and pressures than the hydrogenation of other unsaturated molecules. Similarly, the reaction of benzene with bromine is slow compared with what one expects from an unsaturated hydrocarbon. More significantly, the bromine does not add to one of the carbon-carbon double bonds but ends up replacing a hydrogen atom, yielding bromobenzene (C_6H_5Br).

These examples indicate the kind of special properties that chemists came to associate with benzene and that led them to believe this unsaturated hydrocarbon and others in the same family have unusual chemical properties. That set of properties came to be known as "aromaticity." Perhaps the best way to describe the term is to say that it means that the benzene ring, with its three double bonds, is a unit of surprising stability.

Because of this stability benzene and various derivatives containing the benzene ring are slow to react with hydrogen or other reagents that ordinarily react with unsaturated molecules. The stability also means that when benzene and its derivatives do react, as in the slow reaction of benzene with bromine, there is a tendency for the unsaturated ring structure, with its three double bonds, to be preserved.

These properties were known more than 100 years ago. The structural ex-

planation of them first began to emerge in the 1860's, primarily through the insight of the German chemist August Kekulé. He realized that the structure of benzene I have described is arbitrary, since the decision was made to put double bonds between carbon atoms 1 and 2, 3 and 4, and 5 and 6. An alternative would have been to leave single bonds in those positions and to put the double bonds between carbons 2 and 3, 4 and 5, and 6 and 1. There is thus an ambiguity in the structure, and Kekulé reasoned that this might be related to the special properties of benzene. By his own account he awoke from a dream in which snakes were chasing each other around a circle and realized that the double bonds in benzene could move around the benzene ring so as to spend part of their time between alternate pairs of carbon atoms.

At the time of Kekulé's dream, of course, electrons were unknown. Today we would say that the electrons that form the covalent bonds are delocalized: spread out over the entire benzene ring rather than localized between two particular carbons. Another description is to say that benzene is a "resonance hybrid." That is, it does not have a structure in which double bonds are localized between any particular neighboring pair of carbons but instead has a structure that is a hybrid of the two possible configurations. In this new average structure each carbon has one and a half bonds to each of its neighbor carbons rather than a double bond on one side and a single bond on the other. The same type of structure can be assigned to many other molecules with unsaturated six-member rings, such as naphthalene, pyridine and their derivatives, thus explaining their stability, or aromaticity. When Kekulé proposed that the double bonds in benzene chase themselves around the ring, it was not at all understood how this could stabilize the molecule, but modern quantum mechanics can handle such phenomena quantitatively and readily accounts for the increased stability.

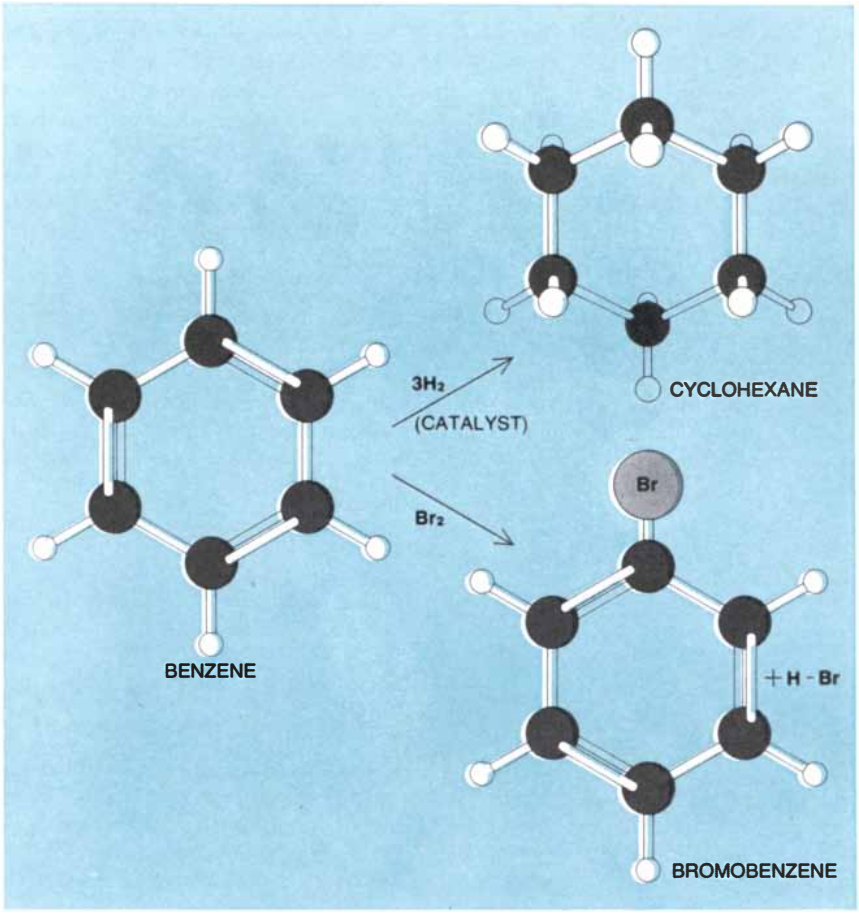
One may ask if it is possible to remove a proton (the nucleus of a hydrogen atom) from a hydrocarbon and leave the hydrogen atom's electron behind in the valence shell of the neighboring carbon atom. It turns out that removing a proton from most hydrocarbons is quite difficult. When it is accomplished, the negatively charged compound that results is called a carbanion.

It has been known for some time, however, that a proton can be removed rather easily from a particular unsaturated hydrocarbon with a five-member ring: 1,3-cyclopentadiene. The resulting carbanion (the cyclopentadienyl anion) resembles a benzene ring from which one carbon atom has been removed, so that the ring contains only two double bonds instead of three. In addition, one of the carbon atoms is drawn with an unshared pair of electrons [*see illustration on next page*].

It is evident that each carbon atom has eight valence electrons, thus possessing a filled outer shell. The reader will also see what chemists realized: It is possible to draw five structures for the cyclopentadienyl anion in which the unshared pair of electrons is successively placed on each of the five carbon atoms. As in benzene itself, the electrons are delocalized, and again it was quickly apprehended that such delocalization is undoubtedly the reason for the anion's unusual stability. As a result the cyclopentadienyl anion can be prepared under conditions far milder than those required for the preparation of other anions.

More recently a positively charged hydrocarbon, a carbonium ion, has been prepared that is also strikingly stable. The compound, the cycloheptatrienylium cation ($C_7H_7^+$), which has seven carbons in a ring and a net positive charge, is remarkably stable for a positively charged molecule. The compound resembles a benzene molecule into which an extra carbon and hydrogen have been inserted [*see top illustration on page 37*].

The seven-member ring contains three double bonds, so that six of the seven carbon atoms have eight valence electrons. The seventh carbon atom is left with only six electrons in its valence shell and hence carries a net positive charge. Again the choice of how to draw the double bonds and where to put the



CHEMICAL REACTIONS OF BENZENE are not like those of a typical unsaturated hydrocarbon. For example, benzene can be hydrogenated to form cyclohexane (C_6H_{12}) by the addition of three molecules of hydrogen (*top right*), but the reaction requires significantly higher temperatures and pressures than the hydrogenation of other unsaturated molecules. Similarly, the reaction of benzene with bromine (*bottom right*) is slow compared with what one expects from an unsaturated hydrocarbon; moreover, the bromine does not add to one of the carbon-carbon bonds but ends up replacing a hydrogen atom, yielding bromobenzene (C_6H_5Br). The unusual chemical properties exhibited by benzene (and similar hydrocarbons) in such reactions are what is described by the term aromaticity.

positive charge is arbitrary; the reader will see, in fact, that seven structures can be written for the cycloheptatrienylium cation. The true structure is a hybrid of the seven, in which each carbon atom has one-seventh of a positive charge and the electrons of the double bonds are spread uniformly around the ring.

Emboldened by examples of this type, chemists set out to prepare new aromatic systems other than those found in nature or already known. One of the earliest surprises came with the synthesis of cyclooctatetraene, which has eight carbons in a ring and four double bonds [*see bottom illustration on opposite page*]. It was expected that this would also be an exceptionally stable, or "aromatic," molecule with the electrons delocalized around the ring.

The actual properties of cyclooctatetraene, when it was finally synthesized by Richard Willstätter in 1913, were not at all like those of benzene. Instead cyclooctatetraene was as reactive as any ordinary unsaturated molecule and showed no sign of any special stability. This finding was so contrary to what was expected that for a long time most chemists believed that the synthesis had been unsuccessful and that a different molecule had actually been prepared.

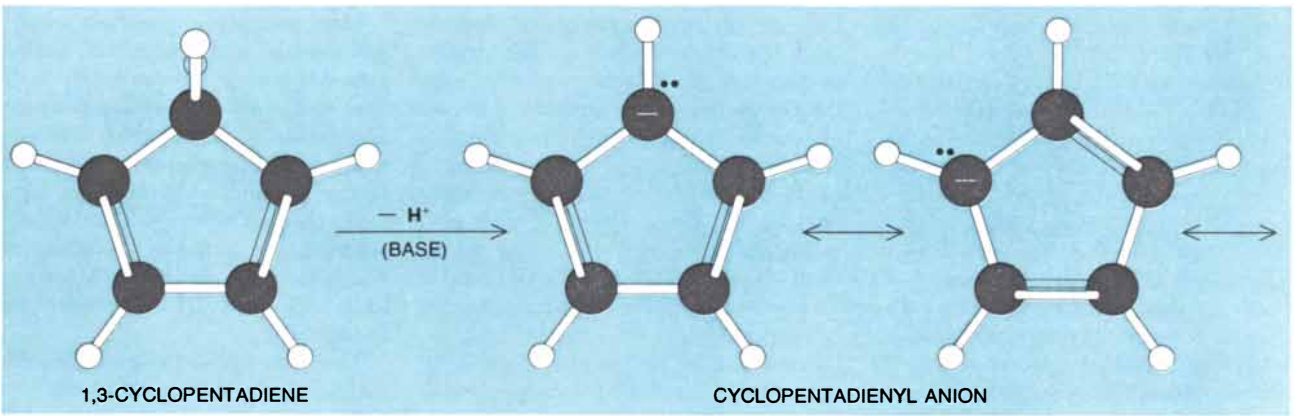
Further surprises awaited those who tried to prepare cyclobutadiene (C_4H_4), a compound with four carbons in a ring containing two double bonds, which could be located in either of two arbitrary positions. Once more it was expected that the real compound would be a hybrid of the two structures and that the molecule would be stabilized, just as benzene is, by electron delocalization. In actuality all attempts to prepare cyclobutadiene by conventional chemical methods led to other materials; the compound proved impossible to prepare. It quickly became apparent that this was because the molecule, far from being strongly stabilized, is extraordinarily reactive. Quite recently it has been possible to generate cyclobutadiene for a fleeting instant in the gaseous state and to observe its reactions with other compounds.

To summarize up to this point, we have seen that an unsaturated six-member ring (benzene) is unusually stable, as is a negatively charged five-member ring (the cyclopentadienyl anion) and a positively charged seven-member ring (the cycloheptatrienylium cation). On the other hand, unsaturated neutral hydrocarbon rings with four members (cyclobutadiene) and eight members (cyclooctatetraene) are notably unstable. Various studies have also shown that positively charged five-member rings and negatively charged seven-member rings are unstable.

Observations of this kind led to the hypothesis that there is something special about the "aromatic sextet." Ignoring the electrons that participate in the single bonds and considering only the electron pairs that get delocalized in the various stable and unstable structures, it is apparent that there are six such electrons in benzene, corresponding to the three electron pairs of the double bonds. (We count only one electron pair in each of the double bonds: the electron pair that can be put elsewhere in an alternative structure.) Therefore in benzene six electrons are delocalized and the others stay essentially fixed between the atoms involved in single bonding. Similarly, in the cyclopentadienyl anion six electrons are delocalized: two electrons for each of the two double bonds and an electron pair that is unshared and leads to the negative charge on one of the carbons.

In the cycloheptatrienylium cation the six electrons of the three double bonds are delocalized. (The positive charge, which is spread around the ring, represents an absence of electron pairs, so that it does not contribute anything more to the total number of electrons delocalized.) In the cyclopentadienyl cation only four electrons are delocalized, and in the cycloheptatrienylium anion eight electrons are delocalized; both configurations are unstable.

What is so magical about the number six? The answer came out of some studies in the 1930's by Erich Hückel on a new way to apply quantum mechanics to molecules, which led to the development of what is called molecular-orbital theory. In this theory the delocalized electrons I have been describing are thought about in a new way. Rather than first assigning electrons to particular locations between pairs of carbon atoms and later deciding that they can be spread out over the entire ring, molecular-orbital theory starts off with the idea that in a system such as a benzene ring there are orbitals, or places where electrons can be put, that spread out over the entire six carbon atoms. Therefore the set of six carbons and six hydrogens of a benzene ring is regarded as a kind of superatom, and electrons are added to it, just as in describing the structure of an atom we consider adding electrons to the atomic nucleus. When this form of molecular quantum mechanics was applied to molecules exhibiting (or lacking) aromaticity, it quickly emerged that



CONVERSION of 1,3-cyclopentadiene, an unsaturated hydrocarbon with a five-member ring, to the aromatic cyclopentadienyl anion can be readily accomplished by the removal of a hydrogen nucleus, or proton (H^+), leaving the hydrogen atom's electron behind in the valence shell of the neighboring carbon atom. The resulting negatively charged compound, called a carbanion, resembles a benzene ring from which one carbon atom has been removed. Five resonance structures are possible for the cyclopentadienyl anion, two of which are shown here. In each case one of the carbon atoms is drawn with an unshared pair of electrons.

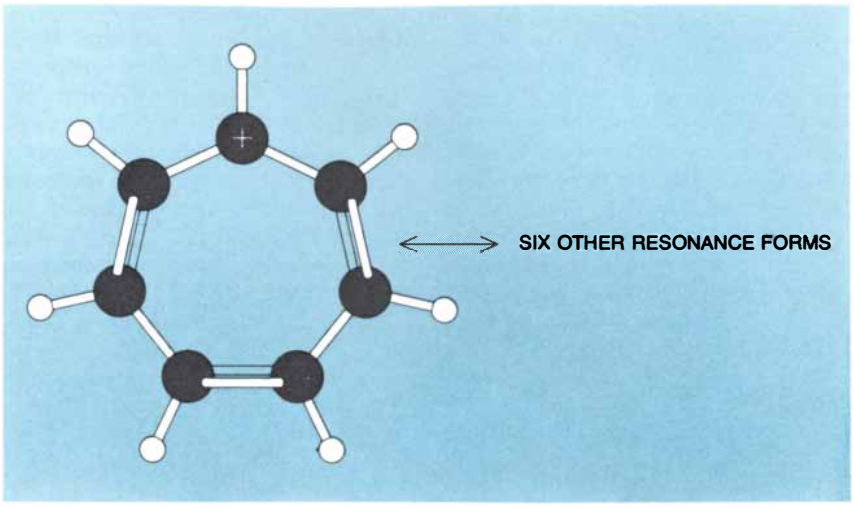
"magic numbers" of delocalized electrons are needed to fill up stable shells, just as in ordinary atoms two electrons are needed to fill a valence shell for hydrogen and eight electrons are needed to fill one for carbon.

According to Hückel's theory, there should be stable delocalized ring systems where the number of electrons needed to fill molecular-orbital shells have the magic values 2, 6, 10, 14 and so on. This has become known as the $4n + 2$ rule, in which n is zero or any integer. The number 6 is the second member of the magic series, and all the aromatic systems discussed up to this point have had six delocalized electrons in a ring.

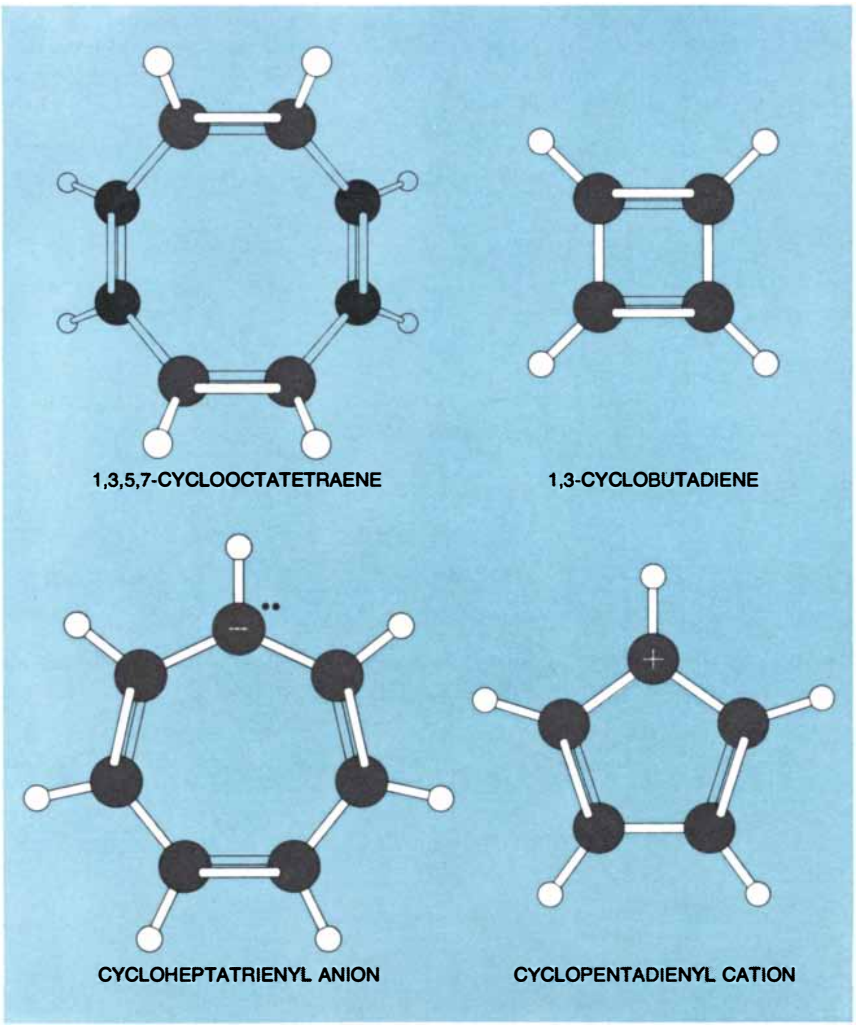
In recent years there has been an intensive effort in many laboratories to explore the implications of this theoretical work. One approach has involved looking for the predicted new aromatic molecules that fit the requirements of the Hückel $4n + 2$ rule. A second important activity has been to try to get further information about the properties of those molecules, such as cyclobutadiene, that do not fit the Hückel rule. Of course, molecular quantum mechanics has progressed since the 1930's, and these experimental observations are guided and interpreted by various new and more sophisticated versions of the original Hückel molecular-orbital theory.

My own work in this area began in 1956 when I arrived as a new instructor in chemistry at Columbia University. I was intrigued by the prediction that there was an aromatic system still to be discovered that would have only two delocalized electrons and so be the simplest system of all. The way to get two delocalized electrons into a ring structure is to construct a three-member ring with one double bond and one positive charge. It would be expected that the electron pair of the double bond would be delocalized over the three carbon-carbon single bonds and that the positive charge would also be spread out over the three carbons. This structure, the cyclopropenyl cation ($C_3H_3^+$), was our synthetic goal [*see top illustration on next page*].

For a variety of technical reasons my colleagues and I concluded that it would be easiest to start off by trying to make a derivative of this system in which the three carbon atoms carried groups other than hydrogen atoms. In 1959 we succeeded in preparing a derivative of the system, incorporating a central cyclopropenyl cation ring, in which each of the three carbon atoms carried what are



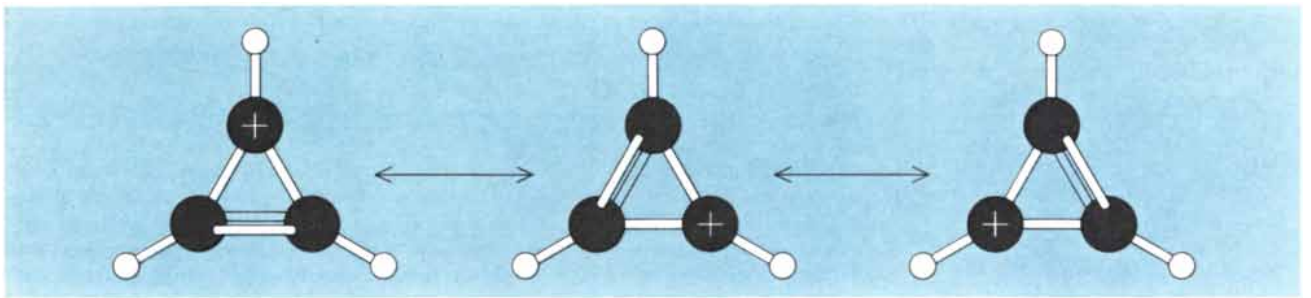
CARBONIUM ION ($C_7H_7^+$), a positively charged aromatic hydrocarbon with remarkable stability, has recently been prepared. The compound, which has seven carbons in a ring and hence is called a cycloheptatrienylium cation, resembles a benzene molecule into which an extra carbon and hydrogen have been inserted. The seven resonance forms differ according to how the double bonds are drawn and where the positive charge is put.



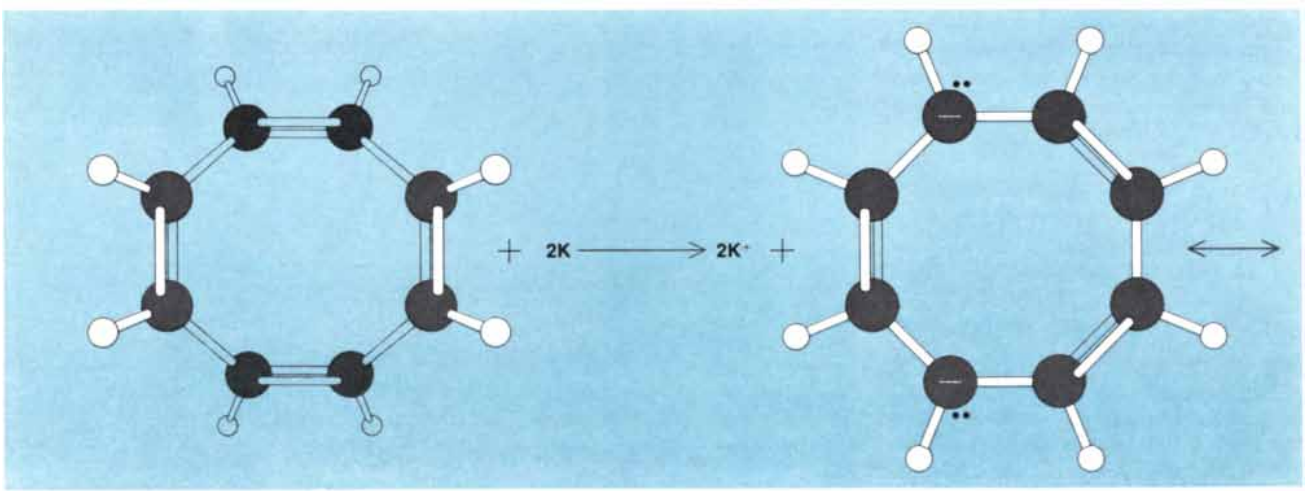
FOUR NONAROMATIC COMPOUNDS, all molecules with the electrons potentially delocalized around a ring, were originally sought in the expectation that they would also be exceptionally stable, or "aromatic." In fact their properties were found to be not at all like those of benzene; all four are at least as reactive as any ordinary unsaturated molecule.

called phenyl groups (C_6H_5) rather than hydrogen atoms. The properties of this molecule, including its remarkable stability, made it quite clear that the central cyclopropenyl cation ring was indeed a new aromatic system. In work extending over the next 10 years our laboratory prepared a variety of derivatives of this simplest aromatic system, whose properties make it fully clear that the cyclopropenyl cation is strongly stabilized by electron delocalization. In 1969, 10 years after we prepared the first derivative of this system, we were finally able to prepare the parent compound $C_3H_3^+$ itself. Its properties were fully consistent with those we had deduced from a study of the other derivatives of the cation and confirmed the prediction that it is indeed the simplest of the aromatic systems.

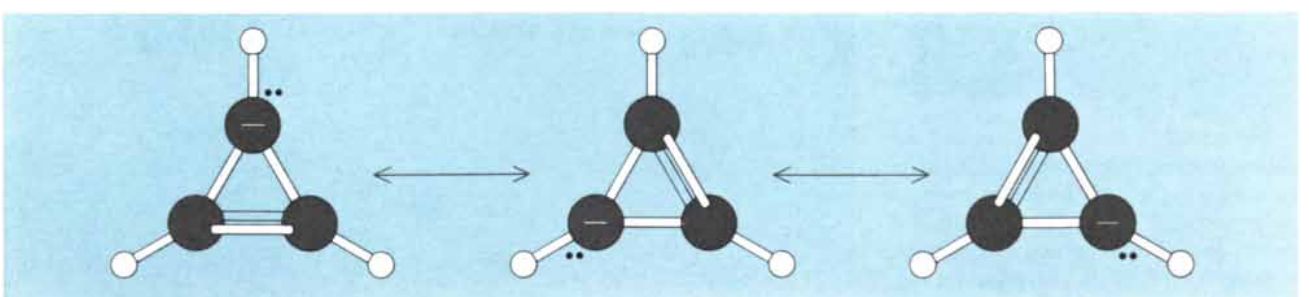
Other chemists have been concerned with extending Hückel's predictions in the other direction. For example, Thomas J. Katz of Columbia has prepared a dianion by adding two more electrons to cyclooctatetraene [*see middle illustration below*]. Whereas the neutral hydrocarbon, with only eight delocalized electrons, was not aromatic, Katz found that the new dianion was in fact an aromatic system with remarkable stability. The two additional electrons fill up the next shell predicted in the molecular-orbital theory, making a system with 10 delocalized electrons in all.



SIMPLEST AROMATIC SYSTEM, the cyclopropenyl cation ($C_3H_3^+$), was first synthesized in 1969 by the author and his colleagues at Columbia University. The two delocalized electrons are spread over a three-member carbon ring that has one double bond and one positive charge. The three possible resonance forms of the molecule are shown. The cyclopropenyl cation and its derivatives (in which the three carbon atoms carry groups other than hydrogen atoms) are all strongly stabilized by electron delocalization.



AROMATIC DIANION ($C_8H_8^{--}$) (*right*) was prepared by Thomas J. Katz of Columbia by adding a pair of extra electrons (provided by two potassium atoms) to the nonaromatic compound cyclooctatetraene, a neutral eight-membered hydrocarbon ring (*left*).

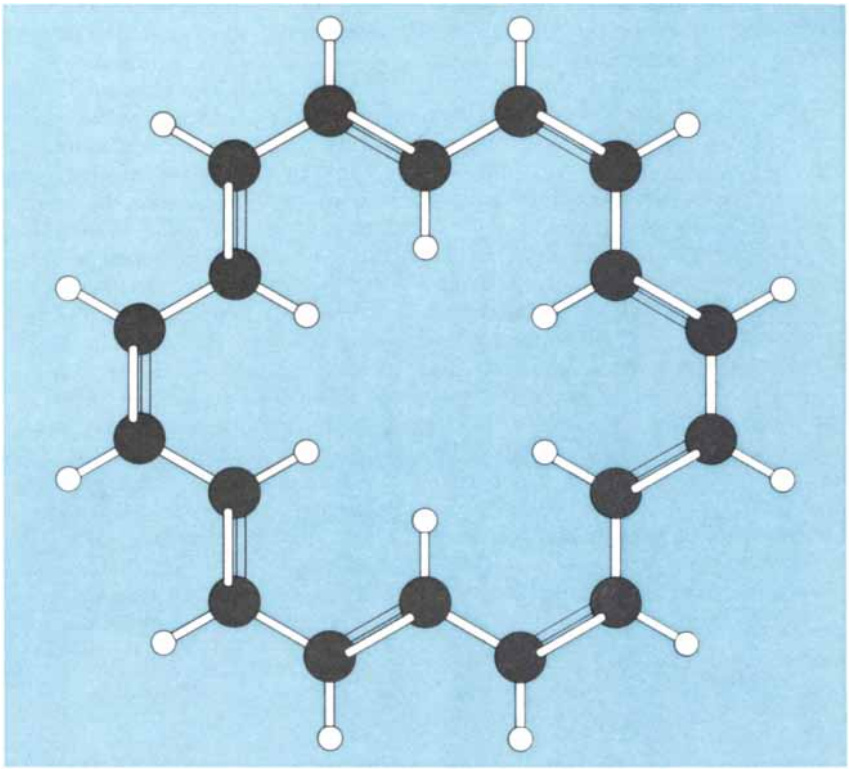


SIMPLEST ANTIAROMATIC SYSTEM, the cyclopropenyl anion ($C_3H_3^-$), is so unstable that it has never been possible to observe its properties directly. Derivatives of the anion have been prepared, however, and from their properties it has been possible to show that the antiaromaticity, or instability, of the system is caused by the delocalization of the "wrong" number of electrons.

Franz Sondheimer of University College London has been working with much larger systems that fit the Hückel predictions. For example, he has prepared a compound with 18 carbons in a ring, each carbon bearing one hydrogen atom [*see illustration at right*]. The nine double bonds in this neutral hydrocarbon can be formally written in two different schemes, indicating that 18 electrons in all are delocalized. The number 18, of course, fits the $4n + 2$ rule.

The properties of this hydrocarbon, which Sondheimer calls [18]annulene, are consistent with its being an aromatic stabilized system. Although the primary interest in these systems was at first simply the attempt to discover a new part of the chemical world predicted by theory, a number of chemists in industry are now exploring the possible applications of the new aromatic systems to the solution of practical problems.

The Hückel rule makes qualitative predictions about 25 percent of the possible delocalized systems one could create but says essentially nothing about what to expect for the other 75 percent, the compounds with $4n$, $4n + 1$ or $4n + 3$ delocalized electrons. Molecular-orbital theory, on the other hand, being a quantitative theory, makes specific predictions about the energies and reactivities of all the delocalized systems. Even the newest forms of molecular-orbital theory, however, contain a number of approximations, and it is not clear how well they do predict the properties of unknown molecules in detail. For this reason there has been considerable interest in exploring the chemistry of the molecules with $4n$ delocalized electrons experimentally to see what kind of properties they have.



LARGE AROMATIC RING, designated [18]annulene, was prepared by Franz Sondheimer of University College London. The nine double bonds in this neutral hydrocarbon can be drawn in two different ways, indicating that 18 electrons in all can be delocalized.

I have already referred to the chemistry of cyclooctatetraene, which appeared to be similar to the chemistry of ordinary molecules with carbon-carbon double bonds. On that basis one might conclude that a molecule of this kind, with eight delocalizable electrons, has normal chemical properties rather than the unusual stability of the aromatic compounds. It is now known, however, that cyclooctatetraene is not a flat eight-membered ring but is strongly puckered. Theory indicates that under these circumstances the electrons cannot be easily delocalized, so that cyclooctatetraene is not a good example of a molecule with $4n$ delocalized electrons.

Perhaps the molecule cyclobutadiene, which I have also mentioned, would be more relevant. From other parts of chemical theory it is quite clear that this molecule should be flat and that the four electrons in the ring should be delocalizable. The observations over the years that cyclobutadiene is an extremely unstable molecule suggested that perhaps the systems with $4n$ delocalized electrons do not have normal properties but are for some reason unstable. One could not exclude, however, a number of other explanations for the high reactivity of cyclobutadiene.

We became interested in this question and decided to investigate the chemistry of the cyclopropenyl anion ($C_3H_3^-$), the negatively charged analogue of the cyclopropenyl cation ($C_3H_3^+$). Like the cation, the anion has one double bond in the three-member ring [*see bottom illustration on opposite page*]. The placement of the double bond, of course, is arbitrary, as is the placement of the negative charge. In the actual delocalized molecule one would expect to find one-third of a negative charge on each of the three carbons, with the electron pair of the double bond similarly spread out among the three carbons.

It is evident that the cyclopropenyl anion, with four delocalized electrons, does not satisfy Hückel's $4n + 2$ rule and thus does not meet the requirements for aromaticity. We hoped to learn whether a system with four delocalized electrons has normal stability or whether the molecule is destabilized by the four delocalized electrons. The prediction that it might be destabilized did not come out of the simple version of molecular-orbital theory originally described by Hückel but was suggested by more recent methods.

It turns out that the cyclopropenyl anion is so extremely unstable that it has never been possible to prepare it and observe its properties directly. In a series of studies extending over the past 10 years we have detected derivatives of the cyclopropenyl anion and have been able to show from their properties that the system is indeed strongly destabilized. Alternative explanations of the instability have been ruled out, so that it seems quite clear that the cyclopropenyl anion is actually destabilized by the delocalization of the four electrons. We have termed this phenomenon antiaromaticity; it reflects unusual instability because of the presence of a delocalized electron system, just as aromaticity reflects unusual stability because of the delocalization of a different number of electrons.

We and others have also studied derivatives of cyclobutadiene and have good evidence that here too the obvious instability of the system is partly due to

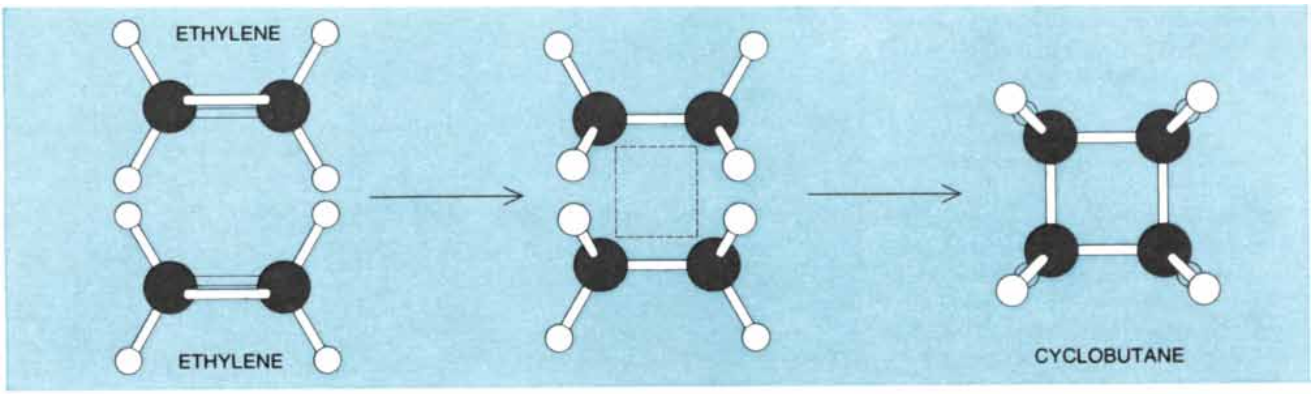
antiaromaticity, reflecting the energetic effect of delocalizing the "wrong" number of electrons. Studies in our laboratory and elsewhere have also found such destabilizing effects for the cyclopentadienyl cation and the cycloheptatrienylium anion, as well as for some other systems with $4n$, rather than $4n + 2$, delocalized electrons.

The fact that in molecules there are magic numbers of electrons that can be delocalized with a gain in stability of the molecule, and other numbers whose delocalization results in instability, has extensive implications. One of the most striking advances involving such concepts is due to Robert B. Woodward of Harvard University and Roald Hoffmann of Cornell University, who recently formulated a very general rule about which chemical reactions can proceed easily and which can proceed only with difficulty.

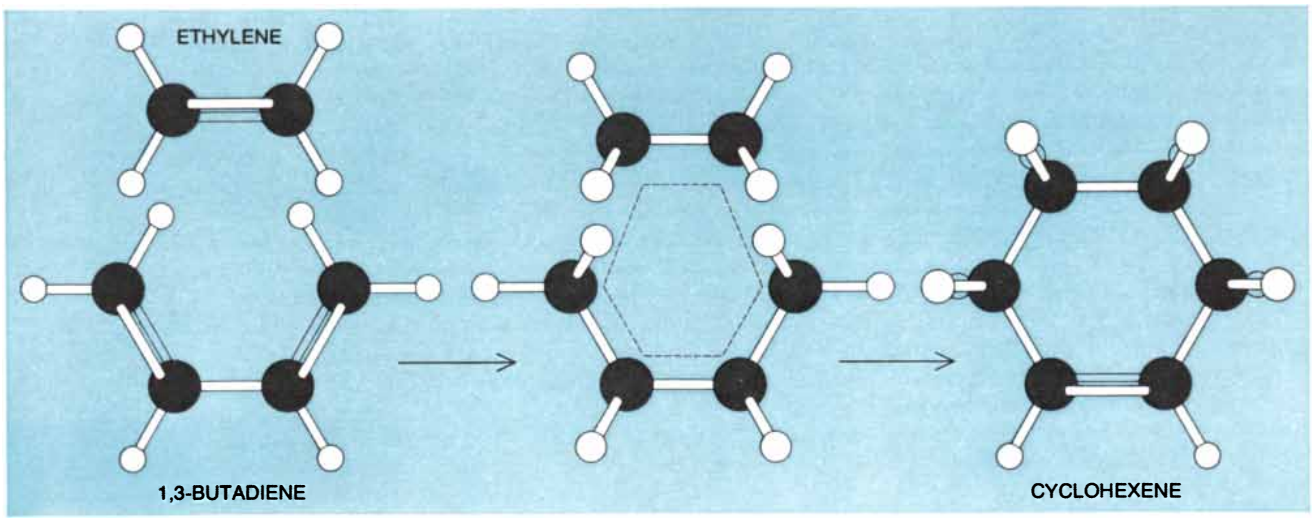
As an example, it is known that two molecules of ethylene (C_2H_4) do not react readily to form the new molecule cyclobutane (C_4H_8) even though the product would be more stable than the starting materials [*see upper illustration below*]. The difficulty is that part of the way through the reaction the two electron pairs of the two double bonds in the two ethylene molecules must be delocalized in order to make the new carbon-carbon single bonds that are required to form cyclobutane. Since four electrons need to be delocalized, the situation is extremely unfavorable.

In contrast, it is relatively easy for ethylene, a molecule with one double bond, to combine with 1,3-butadiene, a molecule with two double bonds [*see lower illustration below*]. Here three double bonds are available to supply six electrons for delocalization as the reaction proceeds, which satisfies the $4n + 2$ rule for stability. Since the work of Woodward and Hoffmann is actually much more explicit and inclusive than these simple examples indicate, it has led chemists to predict reactions that had not previously been observed.

Thus we see that the exploration of aromaticity, and more recently of antiaromaticity, has stimulated important intellectual developments in the theory of molecular structure. At the same time these studies have resulted in the preparation of a number of new types of molecule that are under active exploration as building blocks for future useful materials. Finally, the theoretical advances accompanying this exploration have greatly increased the ability of chemists to predict precisely what chemical reactions will occur and how they will occur, thereby strengthening the entire field of synthetic organic chemistry.



TWO MOLECULES OF ETHYLENE react only with great difficulty to form the new molecule cyclobutane (C_4H_8), even though the product would be more stable than the starting materials, because the reaction must pass through an antiaromatic intermediate state with four delocalized electrons (*broken lines*). In general, ring systems with $4n$ delocalized electrons are strongly destabilized.



ETHYLENE AND 1,3-BUTADIENE, a molecule with two double bonds, can in contrast be combined rather easily to form the new molecule cyclohexene, because the reaction passes through an aromatic intermediate state with six delocalized electrons (*broken lines*). Ring systems whose delocalized electrons satisfy the "magic" $4n + 2$ rule are in general extremely stable and hence are aromatic.

We want to be useful ...and even interesting

Illegitimacy kept him from a notary's career

For two years in Italy, a film about Leonardo da Vinci has been in production. A substantial work, it explores in great depth a life, relentlessly objective of mind and inner-directed at the same time, lived at a point in cultural history before the walls grew up that separate art, science, and engineering. It goes on CBS at 9 p.m. on August 13. Whether you consider photography an art, or part and parcel of science, or a branch of engineering technique, Kodak hopes Leonardo will make five Sunday nights exciting for you.

The woods are full of what? Trees.

These men of the Forest Service are going out in their vehicle "XR-2" to x-ray the trees.

They are of the opinion that x-raying the trees in the woods makes sense. Also city shade trees.

Forty-two years ago the Rochester city forester and a man at the Kodak Research Laboratories brought up the idea in *American Forests*. It did not catch on.

Now there is a different feeling about trees. Cutting trees down just to make sure they are sound no longer seems more sensible than spending money to x-ray them.

Furthermore, it costs less than back in 1930. We now offer sensitized paper for x-rays and a little machine for processing it. No plumbing required. Also, 100-kV x-ray sources have become compact enough to shove easily up a tree.

The men in the picture have taken our course in Rochester on non-destructive testing by x-ray. They have also done a lot to correlate the appearance of the radiographs with forest pests and pathogens. The project is being carried out by the USDA Forest Service Laboratory, P.O. Box 365, Delaware, Ohio 43015.

Non-sentimental note: a big tree can have its dollar value increased by x-ray proof of its internal structure and of freedom from treacherous metal that somebody might have buried in it about the time that Roentgen discovered x-rays.

Kodak

Patient and optimistic

Designs by Drexhage

At Kodak we are no longer discovering laser dyes. Now we design them. Some people design hooked rugs. Karl Drexhage designs laser dyes.

A dye laser is a beautiful light source because not only is it strong and coherent but also tunable. A good dye laser needs a good laser dye. Many kinds of dyes have been tried. We even tried quinine water. One brand sort of worked, and another sort of didn't.

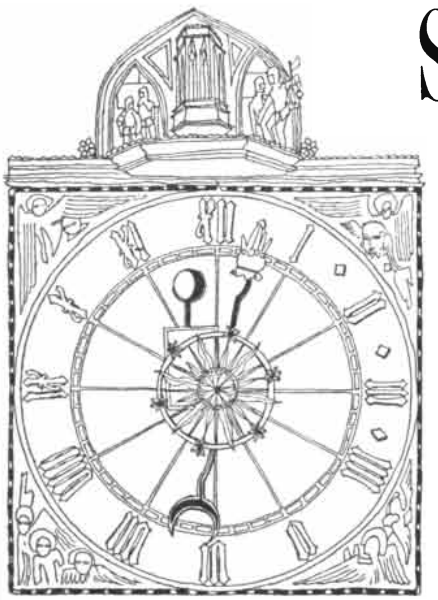
A good laser dye does not lose excitation by electron transfer to or from the molecule. Even that favorite, rhodamine 6G, gets quenched in nonpolar solvents, where the dye salt is poorly dissociated. ClO_4^- and BF_4^- are better than halides as anions. Rhodamine 6G perchlorate keeps its 95% fluorescence quantum yield in 20 very different solvents, reports our Drexhage.

A good laser dye does not lose excitation by conversion to internal vibrations, which is why he makes himself wire models to suggest where steric hindrances could be designed in. They cut down mobility within the chromophore. The electron cloud of a single hydrogen atom can freeze a chemically convenient but physically objectionable substituent rigidly perpendicular to the plane of the actual chromophore. Against losses due to mobility of dialkylamino end-groups, solvents of high dipole moment give localized microrigidity. Another solvent strategem is the use of deuterated alcohol. As the deuterons change places with certain hydrogen atoms on the dye molecule, quantum yield gets better.

A good laser dye molecule resists flipping an electron to enter the condition known as triplet state, where it drops out of laser action. Drexhage has found a rule that predicts how circulating electrons in the molecular structure affect the probability of loss to the triplet state. One of his current favorites for low triplet yield is the phenoxazine class of dyes. With that fundamental advantage, he feels phenoxazines are well worth design efforts to suppress the mechanisms of internal conversion. He is rather pleased with 2,7-bis-(diethylamino)phenoxazine perchlorate in o-dichlorobenzene. Flashlamp-pumped, it lases at 740 nm with a threshold energy comparable to the visible-range performance of rhodamine 6G in ethanol.

Drexhage is not optimistic about efficient dye lasing outside the 0.3–1.5μm range. The limits seem to be set by molecular vibration at the long end and absorption by the excited singlet states at the other end.

Drexhage looks to the future. He would be making a mistake to get tied down by little things like synthesis yield and shelf life that concern the people who make and market EASTMAN *Organic Chemicals and put out "Dyes for Liquid Lasers," which consists of a few sheets packed with data and references for readers who are judged by what they accomplish with laser dyes in the present. Request Kodak Publication JJ-169 from Dept. 412-L, Kodak, Rochester, N.Y. 14650, an equal-opportunity employer.*



The Residue of Stockholm

The 1,200 representatives of 113 nations who attended the long-planned United Nations Conference on the Human Environment in Stockholm in June left behind a mixed record of agreements and disagreements, the ultimate significance of which can only be judged by the individual and collective environmental actions that will be taken in the months and years ahead by the participating nations and by those nations, most notably the U.S.S.R., that did not attend. Three major political and organizational accomplishments, however, did emerge clearly from the two-week parley: (1) a declaration of 26 broadly stated principles, intended to "inspire and guide the peoples of the world in the preservation and enhancement of the human environment," (2) an "action plan" consisting of an assortment of specific recommendations for international measures dealing with problems of pollution and conservation and (3) a proposal for a new environmental coordinating agency within the UN, to be called the Governing Council for Environmental Programmes (GCEP).

The declaration of principles finally approved by the conference follows closely a draft formulated more than two years ago by an intergovernmental working group of the conference's preparatory committee. It enumerates in very general terms a set of "fundamental" human rights and responsibilities with respect to the environment. The prepared draft was challenged successfully at the conference by the Chinese representatives on the grounds that China and some other nations had not had a chance to participate in the original drafting. As a result the draft was reopened to revision and more than 100 amendments were offered and discussed before a specially convened committee representing all the participating nations.

The final document incorporated four new principles—one on the conservation of wildlife, one on the prevention of ocean pollution, one on the problem of helping developing countries to maintain stable prices in spite of increased environmental expenses and one ensuring national autonomy on internal environmental standards. Of the original 23 principles in the first draft, a proposal calling on the nations to give one another information on contemplated actions that would have an environmental impact proved particularly controversial and was referred to the UN General Assembly for further consideration. In addition the language of several other principles was amended in an effort to achieve a consensus. The 26th principle of the declaration, approved only after the president of the conference, Maurice F. Strong, noted that the reservations of China would be taken into account, calls for international efforts to reach prompt agreement "on the elimination and complete destruction" of nuclear weapons and all other means of mass destruction.

As one observer noted, the revised declaration of principles, which was adopted by acclamation at the final session of the conference, amounts to "a sort of constitution against which future environmental actions can be measured by world opinion." The 106 recommendations that comprise the "action plan" also approved by the conference fall into three general categories. The first category of recommendations is aimed at setting up a global environmental-assessment program, called "Earthwatch," in order to "identify and measure environmental problems of international importance and to warn against impending crises." The recommendations in the second category cover environmental management activities and are designed to "put to work what is known or learned about the environment, so as to preserve what is desired and prevent what is

THE CITIZEN

feared." The third category consists of recommendations for various supporting measures, such as education, public information, organization and finance.

The establishment of the 54-nation GCEP, which was praised by Russell E. Train, head of the U.S. delegation, as "the high point of the conference," met some opposition from representatives of other UN agencies, since it will be independent of their control. The new agency, which will administer a voluntary "environment fund" that may eventually amount to some $100 million spread over five years, is designed to coordinate the "action plan" endorsed by the conference. In a final decision the conference recommended that the General Assembly review the entire environmental machinery in 1976.

Alternative Hypothesis

The stage is being set for the mandatory teaching of divine creation as a scientific theory on the same footing as evolution in the public schools of California. In 1969 the State Board of Education modified a new "science framework" for kindergarten through the 12th grade to require that recommended textbooks present more than one hypothesis for the origin of the universe, matter, life and man. The guidelines make it clear that the other hypothesis is to be creation. The advisory committee of scientists and science teachers who drafted the first guidelines objected strongly, but the requirement remains in the science framework.

The author of the revised statement in the science framework is Vernon L. Grose, a vice-president of the Tustin Institute of Technology in Santa Barbara. Grose, who has a B.S. in physics and a master's degree in aerospace management, is also a member of the Curriculum Development and Supplemental Materials Commission appointed by the State Board of Education to evaluate textbooks and workbooks. He favors the teaching of evolution in the schools but feels that divine creation also provides a scientific rationale of first causes. The only accredited scientist on the curriculum commission is Junji Kumamoto, a chemist at the University of California at Riverside. The commission chairman

is Charles Terrell, a school-district superintendent.

The Creation Research Society, a group devoted to the doctrine of special divine creation, has lobbied strongly for the adoption of books that follow the revised guidelines on evolution and creation. Many of the society's members hold master's or doctor's degrees in the natural sciences. Members of the society have been visiting schools in the state to lecture on creationism. A Creation Science Research Center in San Diego has produced a series of study books and handbooks for teachers that present the facts of science in the context of special creation.

Books recommended by the curriculum commission are put on public display for 60 days. A joint meeting of the commission and the State Board of Education is to be held in September or later to formalize the book adoptions. The State Board does not have to accept the recommendations of the commission, but there is the possibility that the board itself may require that the revised science framework be followed. James Ford, a member of the board, has stated that no textbook should be considered for adoption unless it clearly discusses at least the two major contrasting theories of origins, evolution and creation, and that this requirement should be met by more than token acknowledgement.

One California educational publisher has already modified books for the first through sixth grades. Other publishers have been informed that their science submissions for kindergarten through the eighth grade will not be considered unless they contain the discussion on creation theory.

The National Association of Biology Teachers is reportedly considering filing for an injunction if the State Board of Education insists on the requirement. Some local school boards in California also have threatened lawsuits to block the requirement. The injunction would presumably be sought on the grounds that the requirement is unconstitutional as a violation of the First Amendment to the Constitution, which specifies the separation of church and state.

Embattled Think Tanks

The Department of Defense is fighting for the existence of the Rand Corporation, the Institute for Defense Analyses (IDA) and a number of other "Federal Contract Research Centers" that have come under increasing attack from Congress as well as from critics of the military-industrial complex. Last fall, in adopting the military budget for the current fiscal year, Congress cut $8.8 million, or 16 percent, from the amount the Defense Department had requested for the support of four of the 12 nonprofit institutions designated Federal Contract Research Centers. The institutions affected were Rand, IDA, the Center for Naval Analyses (CNA) and the Army's Research Analysis Corporation (RAC). The most heavily cut was IDA, whose Princeton, N.J., office has frequently been picketed by Princeton University students and faculty members and has attracted the adverse attention of some senators because its reports are secret. The IDA budget was cut 23 percent, from $11.2 million to $8.7 million. CNA was cut about 20 percent (to $9.8 million), Rand about 15 percent (to $17.4 million) and RAC 9 percent (to $9 million). Some of Congress' negative attitude toward the "think tanks" was undoubtedly stimulated by the alleged Xeroxing of a copy of the 47-volume "Pentagon Papers" by Daniel Ellsberg when he was a Rand employee. On the other hand, Congress left untouched the budgets of eight other Federal Contract Research Centers, including requests ranging from more than $40 million to $70 million for the Aerospace Corporation, the Applied Physics Laboratory (operated by Johns Hopkins University), the Lincoln Laboratory (operated by the Massachusetts Institute of Technology) and the Mitre Corporation. These four institutions are generally more involved in doing than in thinking.

In making the budget cuts the House Committee on Appropriations stated: "The Committee feels strongly that the time has come for the military services to begin phasing out the 'think tank' operations which have been supported for more than two decades. The level of proficiency and pay in the Government service is such that the Government should be able to move these efforts in-house. The Committee feels that the Government officials responsible for national defense should be more closely involved in these efforts than they are under the present procedures. The Committee further believes that in matters of security better control can be maintained within governmental organizations than outside the Government."

Seeing their budgets growing tighter over the past few years, some of the contract research centers sought to maintain the growth to which they had become accustomed by diversifying their Government sponsorship and doing work for civilian agencies such as the National Science Foundation, the Department of Health, Education, and Welfare and the Department of Housing and Urban Development. For example, in 1969 Rand set up the New York City–Rand Institute to consider problems of fire protection, police, housing and health. Last year about 23 percent of Rand's income came from New York and other civilian sponsors.

In the hope of getting Congress to take a more favorable view of think tanks, the Defense Department has sent John S. Foster, Jr., director of Defense Research and Engineering, and others to testify before Congressional committees. In a recent appearance before a Senate subcommittee Foster explained that the 12 Federal Contract Research Centers receive only 3 percent of the total Defense Department budget for "research, development, testing and evaluation"; another 3 percent goes to other nonprofit institutions. About 29 percent is spent "in the house" and roughly 65 percent in contracts with private industrial and commercial firms. "Having these three different ways of conducting business," said Foster, "provides management options that are desirable and accepted."

Foster classified five of the 12 contract research centers as institutions devoted primarily to studies and analyses. In addition to Rand, IDA, CNA and RAC, the group includes Analytic Services Inc. (ANSER), which advises the Air Force on the cost effectiveness and military worth of most major weapons systems. ANSER's budget is less than $1.8 million. "It has been the purpose of these [five] organizations," said Foster, "to provide a research environment free from the pressures of day-to-day problems.... They have provided us with carefully thought-out 'second looks'—looks which are less susceptible to bias from positions of advocacy. Their hard questions have caused us to face up to issues we might otherwise have missed." Foster added that he was "convinced that the costs of doing work at these organizations are equal to or less than the costs of these same services if performed either in-house or by regular competitive contract."

One important asset of a contract center such as IDA, Foster explained, is a "very healthy" turnover in personnel of from 15 to 20 percent per year (compared with less than 5 percent per year in civil service organizations); in IDA the average professional tour of duty is from five to seven years. "This means," he said, "that a supply of new personnel

with fresh outlooks and recent training is continuously being brought in on defense problems, but under conditions where there is also a large cadre of experienced personnel."

Having conducted this defense of think tanks, Foster listed "a number of corrective actions that I believe we should take to restore [Congressional] confidence in these organizations and in our ability to handle them." Foster promised that the Defense Department would institute closer management controls and review all projects to assure that they cannot be performed effectively and objectively by other organizations. One of the five think tanks, RAC, is being reorganized so that it will no longer have a special relation to the Army. Provided that Congress is willing to supply the funds, the Defense Department would like to support the remaining four study and analysis centers (Rand, IDA, CNA and ANSER) at a constant level, in terms of total manpower, for the next three years. Foster told the Senate subcommittee: "It is obviously hard to contribute effectively to an organization you fear may be closed down tomorrow."

Underdevelopment in the U.S.

It has been known for some time that infant mortality in the U.S. is higher than it is in many other developed countries; indeed, in this respect the U.S. ranks 14th among the nations according to data compiled by the United Nations. A recent survey by the National Center for Health Statistics of the U.S. Public Health Service sheds some light on the matter, finding a strong correlation between infant death and low socioeconomic status of the parents. Moreover, according to the authors, nearly half of the infant deaths in the lowest socioeconomic group are preventable.

The findings are published by the National Center for Health Statistics in a report titled "Infant Mortality Rates: Socioeconomic Factors." The authors are Brian MacMahon and Jacob J. Feldman of the Harvard School of Public Health and Mary Grace Kovar of the National Center for Health Statistics. They write: "Three indexes of socioeconomic status were examined—education of father, education of mother, and family income in the year prior to the birth or the infant death. All three indexes showed a strong association with risk of infant death, this risk being between 50 and 100 percent higher in the lowest socioeconomic class than in the middle and upper classes." A particular finding was that mortality rates were "substantially higher for black than for white infants at all levels of each of the three variables."

The report suggests but does not say directly that the factors underlying the infant's high risk of death in a family of low socioeconomic position are the impoverished environment, the ignorance of the parents about child care and their failure (for lack of knowledge or money) to seek professional help. "Using education of father as the measure of socioeconomic status," the authors write, "and considering as the minimum attainable the mortality rate of 17.4 per 1,000 observed among the infants of fathers in the highest educational class, it can be estimated that 47 percent of the deaths in the lowest socioeconomic group (with a rate of 33 per 1,000) were in excess of this minimum and were therefore, in a broad sense, preventable."

Survival Kit

Some two-thirds of the Americans who die as a result of acute myocardial infarctions (coronary heart attacks) do so before they reach a hospital, many of them in the first hour after the onset of the attack. The incidence of such sudden deaths in the U.S. alone is estimated at more than 300,000 per year, or almost 1,000 per day.

The feasibility of a novel plan to substantially reduce the prehospital mortality rate resulting from these acute heart attacks is now under study at a number of medical-research facilities in this country and abroad, under the general sponsorship of the National Institutes of Health. The essence of the plan is to make available to identifiable "coronary-prone" individuals the means for automatically self-injecting a drug or drugs capable of promptly suppressing the irregularities in cardiac rhythm that are often the prelude to sudden death in the period after an acute heart attack. The basic idea of the plan was originally put forward two years ago by Stanley J. Sarnoff, a former surgeon, medical-school professor and director of research in the Laboratory of Cardiovascular Physiology at the N.I.H.; Sarnoff is now president of Survival Technology, Inc., a Bethesda, Md., company engaged in the development of new medical devices.

As Sarnoff points out in his plan, the immediate precedent of death in many if not most patients who die of coronary heart attacks is ventricular fibrillation, the chaotic and uncoordinated "quivering" contraction of the muscle fibers of the heart's large blood chambers. When a patient goes into ventricular fibrillation in a coronary intensive-care unit in a hospital, he can usually be "defibrillated" promptly by a suitably applied electric shock. When an unattended person outside a hospital suffers ventricular fibrillation, death almost invariably ensues. In most cases ventricular fibrillation is triggered by ectopic, or abnormal, electrical signals originating in the ventricles. The object of the Sarnoff plan is to provide an appropriate procedure for diminishing or abolishing such ectopic beats in coronary-prone patients as soon as possible after the onset of a heart attack.

Drugs that will diminish ectopic heartbeats are readily available. For example, atropine, a drug that increases the heart rate, thereby diminishing ectopic beats, is now frequently administered to heart-attack patients afflicted with bradycardia (an abnormally slow heart rate). In addition to increasing the heart rate, atropine also helps to maintain the blood pressure, which tends to fall dangerously in many heart-attack patients with bradycardia. For heart-attack patients with normal or higher-than-normal heart rates lidocaine (a drug without any influence on the heart rate) is given intravenously almost universally in coronary-care units to suppress ventricular ectopic activity.

What Sarnoff and his colleagues at Survival Technology propose to do is to equip coronary-prone individuals with a "Patient Technology Packet," a kind of cardiac survival kit containing three components, all developed in their company's research laboratory. The components are (1) an "AtroPen," an automatic injector containing atropine, (2) a "LidoPen," an automatic injector containing lidocaine, and (3) a "CardioBeeper," a small electronic device enabling the heart-attack victim to transmit the rate and rhythm of his heartbeat and his electrocardiogram by telephone to his physician or to a central emergency cardiac service. The physician could then have the patient self-administer the appropriate treatment, thereby diminishing the ectopic beats and hence lessening the chance of ventricular fibrillation prior to the patient's arrival at the hospital. The injected drug would also help to maintain cardiac output and arterial pressure while the patient is in transit.

The effectiveness of the AtroPen device for the emergency administration of atropine has already been extensively demonstrated. This particular type of spring-loaded automatic atropine injector was originally designed by Sarnoff

and has been widely adopted by the armed forces of several countries, including the U.S., as a defense in chemical warfare against organophosphorus agents (nerve gases). More recently the same device with the same dose of the same drug has been approved by the World Health Organization as an emergency means of treating intoxication with organophosphorus insecticides. A new-drug application for the use of the AtroPen for this purpose has been filed with the Food and Drug Administration. The increased use of organophosphorus insecticides expected to result from the ban on compounds of the DDT type greatly enhances the need for an automatic atropine injector such as the AtroPen in this field.

A cooperative study designed to develop information about the effects of atropine and lidocaine when administered with automatic injectors early in acute heart attack was recently launched by the National Heart and Lung Institute. The participating laboratories include, besides Survival Technology, the Yale University School of Medicine, the University of Pennsylvania School of Medicine, the Binnengathuis Hospital in Amsterdam, the Nassau County (N.Y.) Medical Center and the Montgomery County (Md.) Heartmobile. Meanwhile studies are under way elsewhere to document the role of the CardioBeeper in providing accurate and reliable telephoned information on the basis of which a physician can confidently prescribe the correct drug for the heart-attack victim to self-administer.

F.D.A. approval must be received before the Survival Technology plan can be implemented. Nonetheless, Sarnoff and his colleagues believe their plan is "fundamentally sound" and report that they are increasingly confident that "the time is approaching when we will have the opportunity to participate effectively in a program that will save many human lives."

Petite Différence

Since male sex hormones are most abundantly produced by males and female sex hormones most abundantly produced by females, one might suppose that the sexual receptivity of each sex was mediated by the predominant sex hormones. Indeed, that is the case with most mammals. It is not, however, with primates, including man. This paradoxical finding has been made by B. J. Everitt and his colleagues at the University of Birmingham in studies of the sexual behavior of rhesus monkeys.

Earlier work by the Birmingham group established the fact that removal of a female monkey's ovaries (thereby eliminating its supply of estrogen and progesterone, the female sex hormones) did not reduce the female's sex drive, although it markedly depressed the sexual behavior of the males with whom the ovariectomized female was paired. It has also been observed that removal of the ovaries does not reduce the libido of human females. Recently Everitt and his colleagues have learned that when the female monkey's adrenals—the principal secretors of the male hormone androstenedione—are removed, her sexual receptivity is greatly reduced.

The investigators studied females, all without ovaries, whose normal sex-hormone balance was maintained with injections of estrogen. Some of the subjects also had their adrenal glands removed. The females were divided into three experimental groups. The first received no hormone supplement except estrogen. The second received both estrogen and androstenedione. The third received both estrogen and a second secretion of the adrenal gland, dehydroepiandrosterone. Each female was then paired once a day with the same male monkey over a period of a month.

The female control group, with adrenals intact, showed a normal degree of sexual receptivity and initiated a majority of all the sexual contacts. The females without adrenals who received estrogen alone or together with androstenedione showed greatly reduced receptivity. The group that received both estrogen and androstenedione behaved very much like the group with intact adrenals.

Everitt and his co-workers conclude that whereas in female nonprimates the secretion of female sex hormones usually enhances sexual responsiveness, among female primates the male sex hormones, secreted partly by the ovaries but mainly by the adrenal glands, have taken over this function. Although the female sex hormones continue to play an important role in primate sexuality, they do so by increasing the sexual attractiveness of the female; hence their principal effect is to stimulate the male. The female's male sex hormones come into play in maintaining and amplifying her responsiveness to the male's advances.

Pure and Simple

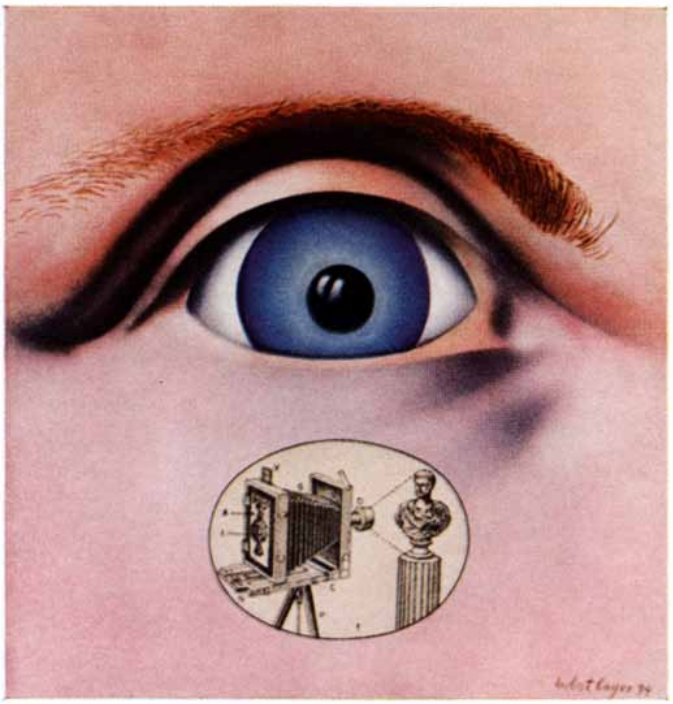
Hydrogen, which has many uses in industry and the laboratory, is not easy to store. It can be held either as a liquid below the temperature of 20 degrees Kelvin or as a highly compressed gas, but neither procedure is economical or particularly safe. Workers at the Philips Research Laboratories in the Netherlands may have found another way: they have successfully synthesized intermetallic compounds in which the hydrogen is held in the form of hydrides. The hydrogen can be released and reabsorbed at room temperature and at pressures of a few atmospheres.

Compounds of a rare earth and nickel or cobalt are capable of binding large amounts of hydrogen. Basically what happens with the nickel compounds, such as $LaNi_5$ (lanthanum and nickel), is the following. If at least six hydrogen atoms are incorporated per unit of $LaNi_5$, a new chemical compound is formed: a hydride. At a given temperature and pressure the $LaNi_5$ and the hydride are in equilibrium with each other. The amount of hydrogen in the mixture can be varied within wide limits, and the equilibrium pressure has a value that is practically constant (called the plateau pressure). If the $LaNi_5$ is put in contact with hydrogen at a hydrogen pressure slightly above the equilibrium pressure and the temperature is dropped to compensate for the heat generated by the reaction, the hydrogen is absorbed by the $LaNi_5$ until only the hydrogen-rich hydride is present. If hydrogen is now allowed to escape, the pressure falls rapidly to the plateau pressure and remains there until all the hydrogen is drained away. While the hydrogen is escaping the hydrogen-rich hydride is reconverted to the hydrogen-poor $LaNi_5$. Only after the process is complete will the pressure drop below the plateau pressure. In the cobalt compounds the situation is slightly more complicated because at a certain temperature and pressure different hydrogen-rich compounds can be formed.

The process of storing hydrogen in hydrides has several attractive features. For example, at a pressure of four atmospheres hydrogen in $LaNi_5$ has a density that in the absence of $LaNi_5$ could only be achieved at 1,000 atmospheres. The storage is quite safe, since during the liberation of hydrogen from $LaNi_5$ the temperature drops. If a storage vessel were to leak, the flow of hydrogen would stop automatically because the speed at which the gas is released decreases with a drop in temperature. Moreover, because the intermetallic storage compounds absorb hydrogen selectively, they can also be used to purify hydrogen gas.

The ideal

We must teach our children not only what to see but *how* to see.



artist: herbert bayer (1934)

The real

Too often we try to educate by preaching orthodoxies. But orthodox solutions are no longer enough.

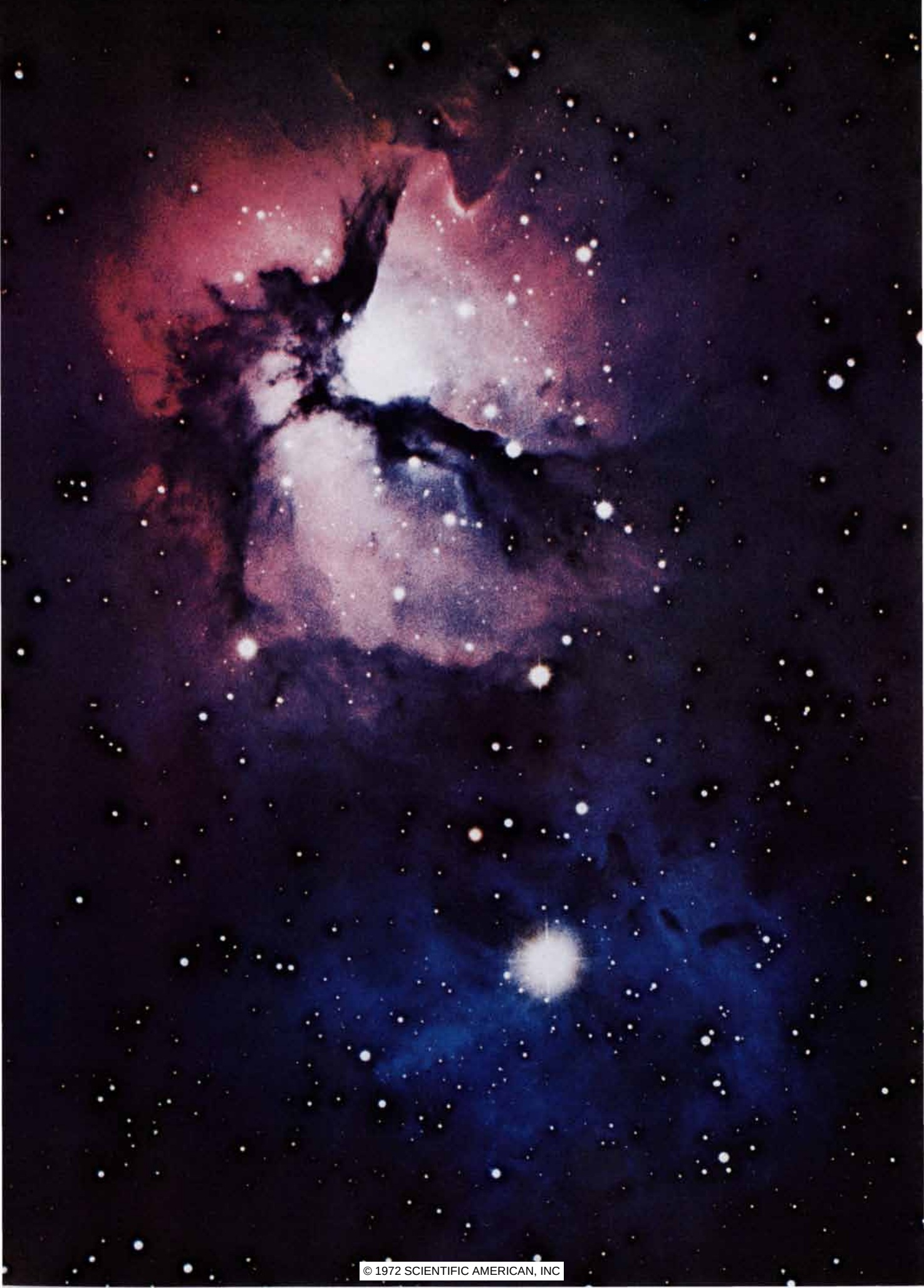
Our world is changing faster than ever before while many of our educational premises remain static, mired in the past.

But the past no longer has enough of the answers. In the years ahead, problems will arise for which there are no precedents. To keep the future open we must teach our children not only what to learn but *how* to learn, *how* to see, *how* to analyze.

Only then will they be able to recognize and cope with problems which our generation cannot even foresee.

AtlanticRichfieldCompany

artist: jasper johns



THE BIRTH OF STARS

Since stars grow old and die, they are presumably born. In our galaxy the most favorable conditions for their birth are found in the clouds of dust and gas along the central galactic plane

by Bart J. Bok

It is clear that the stars in our galaxy are of many different ages. The great majority of them, such as the sun, are old by cosmic standards and will live for some billions of years. At the end of their life many of them will subside to being white dwarfs: tiny stars about the size of the earth that will eventually cool to a dark cinder. There are other stars, however, that are consuming their nuclear fuel so rapidly that they cannot have existed as stars for more than a few million years; they include the hot blue-white giants such as Rigel in the constellation Orion. They must evolve rapidly through the gradual exhaustion of their internal energy sources and become less conspicuous. If stars have finite lifetimes and can grow old and die, it would seem that they must also be born. What is the evidence for the birth of stars in the galaxy?

The visible band of the Milky Way marks an almost perfect great circle in the sky, indicating that the sun and its planets are close to the central plane of the galaxy. Counts of the stars along the Milky Way show that the fainter (and therefore the more distant) stars are much more concentrated toward the galactic equator than the brighter ones, which suggests that the Milky Way system is a large galaxy that is highly flattened. The uneven distribution of faint stars along the Milky Way shows that the sun is not located at the center of the galaxy; modern estimates place the sun at a distance of 30,000 to 33,000 light-years from the center. The direction toward the center is marked by the great star clouds in the constellation Sagittarius.

The main body of the galaxy is a disk with an overall diameter of some 100,000 light-years. The flatness of the disk suggests that it rotates at a fairly rapid rate, and that is the case. It does not, however, rotate like a solid wheel. The inner parts complete a circuit around the center much more quickly than the outer parts do. The observational evidence for galactic rotation is that stars in the neighborhood of the sun generally move in almost circular orbits around the galactic center at an average rate close to 250 kilometers per second. The distance from the sun to the galactic center is so great that it takes the sun 250 million years to complete a single circuit; we may call that interval a cosmic year. The principal gravitational force that controls this motion is produced by the central star clouds of the galaxy, which have a total mass of perhaps 50 billion suns. The mass of the entire Milky Way system is estimated to be a little more than 100 billion solar masses.

The appropriate conditions for the formation of stars exist only in the central plane of the galaxy or close to it. The processes of star birth have apparently exhausted themselves in the spheroidal "halo" surrounding the galaxy, where old stars prevail. These processes, however, are very much under way in the central disk, where interstellar gas and dust intermingle with young stars. It is believed that the interstellar gas and dust are the material from which stars are now being formed. Dense concentrations of gas and dust that are on the way to becoming real stars are termed protostars. It has been suggested that protostars are formed when some of the gas and dust associated with the spiral arms of the galaxy piles up into clouds. A more specific possibility is that observed small dark clouds of interstellar dust grains and even smaller dark "globules" of dust collapse into stars or clusters of stars. A third suggestion is that protostars originate with gas ejected in the explosions of supernovas; a few such clouds are visible as luminous filaments. We shall be considering each of these processes in turn.

The stars in the galaxy are being continually reshuffled. In the course of one million years—only a two-hundred-and-fiftieth of a cosmic year—two stars now close together but having a difference in velocity of one kilometer per second will have separated by three light-years. Therefore in less than one cosmic year some groups of stars may be dissipated and others may be formed afresh; the appearance of the galaxy should show profound changes. The physical makeup of the Milky Way system will also change over intervals of the order of one cosmic year. As we have seen, the rate at which some kinds of stars use up their supplies of energy is quite high. The blue-white Type O and Type B giant stars are a good case in point. They consume their supplies of nuclear fuel at such a prodigious rate that they cannot keep it up for more than a fraction of a cosmic year. The majority of these stars have probably existed 10 million years since their formation, which is only a twenty-fifth of a cosmic year.

The loosely bound "open" star clusters

TRIFID NEBULA, No. 20 in the Messier catalogue of celestial objects, is a good example of a bright cloud of interstellar gas associated with a dark cloud of cosmic grains. The Trifid Nebula is in the plane of the Milky Way; there is evidence that some of its material is condensing into protostars, dense concentrations of gas and dust that are on the way to becoming real stars. The blue nebulosity at the bottom right has been excited to a very high temperature by the central bright star; the red nebula above is somewhat cooler. Color photograph was made with the 200-inch Hale reflecting telescope on Palomar Mountain.

are the best indicators that star formation is still a continuous process. Within a couple of thousand light-years of the sun there are a good many open clusters with luminous Type O and Type B stars. Open clusters are groups of several hundred stars within a volume roughly 10 light-years in diameter. The presence of Type O and Type B stars implies that many such clusters must be quite young on the cosmic time scale. If open clusters evolve physically, each of them should pass through a rather brief period of glory and then subside and become a less spectacular grouping of old stars.

In the first stage of the cluster's development the protostars will evolve and move toward the "main sequence" of the Hertzsprung-Russell diagram [*see illustration on page 52*]. Most of the stars will spend their early life on the main sequence, where they will use up their principal supplies of nuclear energy. These supplies are exhausted at the highest rate by the intrinsically bright stars, which are the first to evolve away from the main sequence. Most of them will shed much of the gas in their atmosphere and evolve toward the white-dwarf stage. Stars that have an intrinsic brightness less than or equal to the sun's will spend a few billion years on the main sequence. An evolved star cluster will have very few highly luminous stars and a number of white dwarfs.

If, as seems likely, the formation of loose clusters with Type O and Type B stars has been taking place at a reasonably steady rate over the past 10 to 20 cosmic years, one should expect to find 100 old loose clusters for every spectacular young one. This is not the case. Where have the clusters gone? The scarcity of old open clusters probably indicates that they evolve rapidly and dynamically. Relatively straightforward calculations show that clusters with a few hundred members and with diameters of the order of 10 light-years lead a precarious existence in the galaxy. The nucleus of the galaxy exerts tidal forces on them that are highly disruptive. The individual stars are gravitationally so loosely bound to the cluster system that quite a few of them, particularly the less massive ones (which move the fastest), will escape from the cluster in a cosmic year or so. In addition, encounters and near-passages between star clusters and clouds of interstellar dust and gas will tend to loosen up most open clusters and disrupt them in less than a cosmic year.

What are the processes that allow the supply of young open clusters to be maintained? There are several places where the birth of clusters and stars may be in progress. The Pleiades and Hyades clusters in the constellation Taurus will almost surely not be there a cosmic year in the future, but replacements seem to be on the way.

To understand the processes of star formation more fully, it is necessary to ask first how the galaxy was formed. The oldest star clusters and individual stars are found at large distances from the central plane of the Milky Way. This fact would seem to imply that shortly

LANE OF DUST AND GAS is plainly visible in the plane of the spiral galaxy Messier 104 in Virgo, which was photographed with the 200-inch Palomar telescope. The galaxy is surrounded by a spherical "halo" of mostly old red stars, and star birth is confined to the galactic plane. The Milky Way galaxy, if it were viewed from the outside, would probably look quite similar to Messier 104.

after the universe was formed the Milky Way system became a separate unit, a large, nearly spherical blob of gas. That may have happened 40 or 50 cosmic years ago. When the first condensation began to form in the original gas and dust, stars and star clusters were probably born all through the large blob of gas. Globular star clusters, which are much more symmetrical and much richer in stars than open clusters, were apparently formed between 20 and 40 cosmic years ago, suggesting that conditions in the original gas cloud were relatively quiescent. As time progressed the gas began to be concentrated more toward the central plane of the galaxy, where it somehow achieved its present rotational properties. Younger stars and clusters were formed in the gas cloud as the cloud became increasingly more flattened.

The galaxy is currently in a stage of development where its central gas and dust layer is remarkably thin: only 1,000 light-years thick in the vicinity of the sun. At present star birth seems entirely confined to this thin layer of gas and dust. It is fortunate for astronomers interested in the birth of stars that the evolutionary processes are continuing near the central plane of the galaxy and even more fortunate that the sun and the earth occupy a position well suited for the observation of these processes.

The Interstellar Medium

In order to get a clear picture of star formation and protostars one must consider the physical conditions in the interstellar medium. It is in this medium that the concentrations that give rise to protostars are formed. The composition and physics of the interstellar medium are now quite well understood. The principal constituent is hydrogen. The ionized hydrogen atom can be detected in a great variety of ways. In the visible portion of the spectrum it announces its presence by the Balmer recombination lines observed after a free electron is captured by a positively charged hydrogen nucleus (a proton). The Balmer lines are emitted as the electron cascades to the second level of the neutral (un-ionized) atom [*see illustration on page 53*]. Radio astronomers catch the hydrogen atom in transitions between very high energy levels, or they observe the continuous spectrum of radiation generated as an electron whizzes past a proton. Neutral atomic hydrogen emits and absorbs radiation at the radio wavelength of 21 centimeters and can be readily observed by radio telescopes. Molecular hydrogen has been detected by means of ultraviolet observations from rockets. It seems only a matter of time before its distribution will be charted in detail through ultraviolet observations, possibly combined with infrared and radio observations.

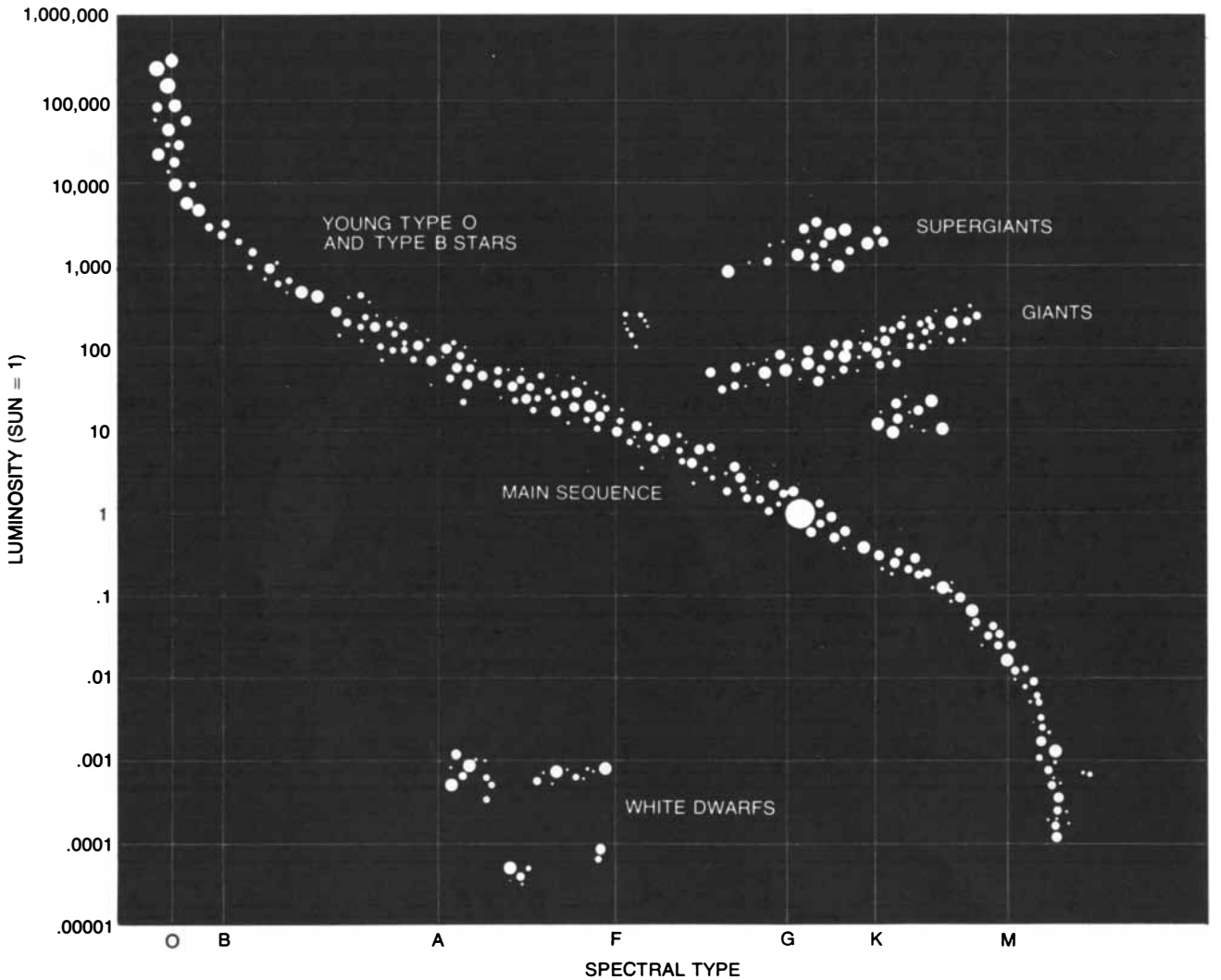
Also present in interstellar space are helium, nitrogen, carbon, oxygen and many other elements. The spectral lines emitted by these atoms and their ions are best observed in the brightly glow-

YOUNG STARS AND A LARGE NEBULA are found side by side in the region of Messier 8, photographed by the author with the 90-inch reflecting telescope at the Steward Observatory of the University of Arizona. It seems as though the processes of star birth have been more or less completed in the cluster on the left, and dust and gas are apparently condensing in the nebula on the right.

ing emission nebulas. When any gas is heated, it will emit radiation at certain wavelengths, depending on its composition. If the gas is cooler than a star behind it, it will absorb the star's radiation at those same wavelengths. The elements present in the interstellar medium are checked whenever possible by the interstellar absorption lines observable in the spectra of distant stars. For every 10,000 hydrogen atoms there are on the average 1,200 helium atoms, two carbon atoms, one or two nitrogen atoms, three or four oxygen atoms, one neon atom, one sulfur atom and traces of some of the heavier atoms such as iron and chlorine. If a laboratory chemist were able to analyze a sample of the interstellar gas, he would conclude that it was a mixture of hydrogen and helium with impurities!

Not only atoms and ions but also molecules have been detected in interstellar space. In the 1930's some of the telltale absorption lines found in the spectra of distant stars were attributed to simple diatomic molecules such as CH, CH^+ and CN. The past 10 years have witnessed an amazing increase in the number of molecules detected. In 1963 lines in the radio spectrum were discovered that were attributed to the hydroxyl radical (OH). Three major discoveries followed in 1968 and 1969: molecules of ammonia (NH_3), water vapor (H_2O) and formaldehyde (H_2CO). In 1970 a number of additional lines were discovered, notably one for carbon monoxide (CO). It is striking that some very complex molecules, for example methyl alcohol (CH_3OH), are present. The list of interstellar molecules has now been extended to some 30. Most of the molecules are found in regions of space where cosmic dust prevails. The Great Nebula in Orion, the gaseous clouds near the center of the galaxy and the quiescent large globules and larger clouds of cosmic dust within a few hundred light-years of the sun have proved to be the favorite hunting grounds, but molecules continue to pop up in surprising concentrations in many unexpected spots. They are often found in regions near sources emitting strong infrared radiation.

There has been much discussion in recent years about the composition and physics of cosmic dust grains. Their pres-

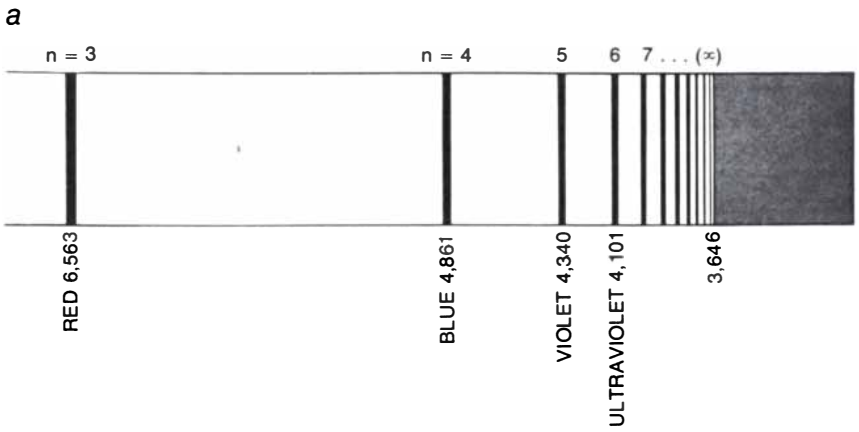


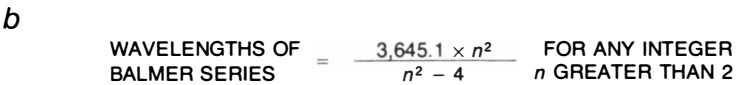
HERTZSPRUNG-RUSSELL DIAGRAM relates the luminosities and the spectral classes of stars. Luminosity in terms of the sun is plotted along the vertical axis and spectral classes along the horizontal axis. The spectral type corresponds to temperature and color. Type O and Type B stars such as Spica in Virgo are bluish white and have a temperature of about 20,000 degrees Kelvin or higher; Type M stars such as Antares in Scorpio are red and have a temperature of about 3,000 degrees K. The hot young Type O and Type B stars typical of the spiral arms of galaxies occupy the top left corner of the diagram. They are extraordinarily luminous and are consuming their nuclear fuel at such a prodigious rate that they cannot have existed in their present state for more than 10 million years. Most older and more "normal" stars such as the sun (*large dot*) are grouped along a band called the main sequence that stretches from the top left corner to the lower right corner of the diagram. These main-sequence stars have a lifetime of billions of years.

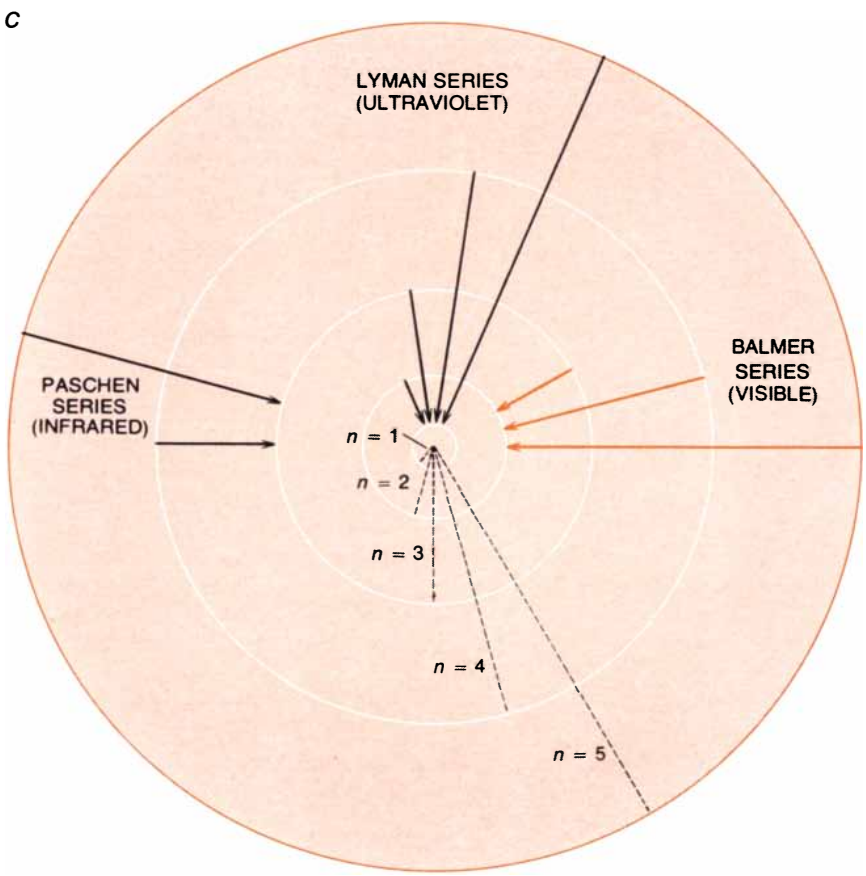
ence in interstellar space is conclusively demonstrated by the existence of dark nebulas and globules of all kinds and by the reddening of starlight as the dust grains scatter the bluer light. Stars close to the galactic equator are reddened, showing that the central plane of the galaxy is rich in dust grains. The infrared radiation observed from many objects in the galaxy seems to indicate that stars or protostars are embedded in thick clouds of dust grains. The grains are tiny particles with a diameter of the order of .0005 millimeter. At first it seemed that they were specks of "dirty ice" built of simple molecules of carbon, nitrogen and oxygen combined with hydrogen and possibly contaminated with iron and other substances. That interpretation was abandoned when observations in the infrared failed to yield evidence for frozen water. Cosmic dust grains were also found in considerable abundance in regions hot enough for "dirty ice" to have evaporated. Grains of graphite with an icy mantle were suggested, and for a while the hypothesis held much favor. Next it was found from studies in the infrared that silicate particles are probably abundant in the atmospheres of very cool stars rich in oxygen; it seemed not unlikely that many of these particles would be lost to the interstellar medium. Present conceptions are that the interstellar grains are most likely a mixture of graphite, silicate and iron grains. They probably originate mostly in the extended atmospheres of infrared stars and starlike objects. It is not out of the question that some grains have been ejected into interstellar space as a by-product of supernova explosions. The trend is definitely away from supposing that cosmic dust grains were formed to any great extent in the cool interstellar clouds.

The neutral atomic hydrogen in the spiral arms of the galaxy is probably at fairly low temperatures: about 100 degrees Kelvin (degrees Celsius above absolute zero). The gas in the regions between the arms is in all likelihood at a far higher temperature, possibly as high as 10,000 degrees K. It is more rarefied than the gas in the spiral arms, but because of its higher temperature there may be a pressure equilibrium between the gas in the spiral arms and the gas in the regions between the spiral arms.

What are the conditions inside the clouds of cosmic dust grains: the dark nebulas? Radio studies of formaldehyde and other molecules found in dark clouds have shown that the temperatures inside these clouds may be as low as five degrees K., and in some places they are probably even lower. There seems to be







BALMER LINES in the visible portion of the hydrogen spectrum (*a*) indicate that the principal constituent of the interstellar medium is hydrogen. The lines are produced when the single electron of the hydrogen atom cascades down to the second energy level of the atom. The exact wavelength of the line is specified by a simple formula (*b*). If the electron drops from the third energy level ($n = 3$) to the second ($n = 2$), it emits light at the wavelength of 6,563 angstroms (where one angstrom is 10^{-8} centimeter). If a free electron is captured from infinity ($n = \infty$) by a positively charged hydrogen nucleus and falls to the energy level $n = 2$, the wavelength of light emitted is 3,646 angstroms. Similar series of spectral lines are produced when the electron drops to other energy levels of the atom (*c*).

little doubt that conditions deep inside large dark nebulas and probably in globules are right for the formation of many kinds of molecules. Gaseous atoms should often stick to the very cold solid interstellar grains inside these clouds. The small solid grains therefore provide surfaces where atoms can combine to form molecules. They also serve to filter out most of the ultraviolet radiation that might penetrate inside the dark cloud or globule. The key to the formation of interstellar molecules seems to lie in low temperature and in the absence of ultraviolet radiation, which would inhibit the formation of molecules and would also destroy the molecules that had formed.

Many kinds of dark nebulas have all the necessary properties for becoming protostars. In a given nebula counts of the number of stars seen through it and studies of their colors yield information about the cloud's absorption of light and its probable distance from the sun. The cloud's linear diameter and its mass in terms of the cosmic dust can then be estimated. If the spectral lines of molecules are found within the cloud, they tell of the composition and mass of the gas associated with the dust. The relative intensities of certain molecular spectral lines and the presence or absence of others indicate the order of magnitude of the temperatures inside these prestellar concentrations of dust. Turbulence within the clouds widens the spectral lines, giving information about turbulent velocities inside the clouds. The turbulent velocities are quite low, often only a few kilometers per second. All these properties must be known before a theoretician can try to construct a hypothetical model of a protostar.

Star Formation and Spiral Arms

Spiral galaxies rank among the grandest of celestial spectacles. They have been studied intensively in recent years with both optical and radio techniques. The spiral arms and their related features, which show up magnificently in photographs of neighboring galaxies, have been subjected to intense scrutiny, and their composition and dynamics are beginning to be understood. The Milky Way has the same kind of spiral structure. Work on it and neighboring galaxies has shown that the spiral arms are concentrations of interstellar gas and cosmic dust of more than average density associated with hot Type O and Type B stars. It is significant that the young Type O and Type B stars are found, singly or in clusters, along the spiral arms; they show up in photographs as brilliant beads lighting up the string of the spiral arm. These young stars generate the ultraviolet radiation that produces the beautiful bright emission nebulas characteristic of spiral features. Radio studies show that the interstellar hydrogen and other gases are distributed more densely along the spiral features than between them. Dust is found throughout the arms, and as one would expect it is associated with the interstellar gas. The highest dust concentrations are found mainly along the inner parts of the arms closest to the central nucleus of the spiral galaxy.

The stars commonly found in the spiral features are without exception very young. They are between 10 and 25 million years old, ages between a quarter and a half of 1 percent of the age of the sun and the earth, or at most a tenth of a cosmic year. It is therefore only natural to look for evidence of continuing star birth as a phenomenon of spiral structure. Are the physical conditions in spiral features such that the formation of protostars will naturally take place? How will the protostars, once they are formed, collapse into the dense nebulous objects that are observed by their infrared emission? How will these objects in turn develop into the hot Type O and Type B stars and other young stars that are found preferentially in the spiral features?

Theoretical work on the spiral structure of both the Milky Way and external galaxies provides good evidence that phenomena of shock and compression are responsible for such features. The theory most widely accepted for the production and maintenance of large-scale spiral structure is the density-wave theory developed by C. C. Lin of the Massachusetts Institute of Technology and Frank H. Shu of the State University of New York at Stony Brook. The theory predicts that a gravitational potential wave of spiral form rotates within each spiral galaxy, moving like a boomerang in the central galactic plane. When the density wave passes through the interstellar medium, dense concentrations of dust and gas are produced, which become detectable as spiral arms. W. W. Roberts of the University of Virginia has shown that a high-pressure shock wave accompanies the density wave. If clouds of cosmic dust and gas of more than average density are present in the interstellar material, they are compressed to five or 10 times their original density—possibly beyond the critical stage necessary to form protostars. The shock wave may serve as a trigger mechanism that sets off the process of star birth along a spiral line. Conditions for the piling up of dust and gas are most favorable along the inside of the spiral arms. Here is where the early beginnings of star formation are found. The entire process takes place within a short time: an interval of the order of one million to 10 million years. One would expect to find the newborn stars along the inner edge of a spiral arm. Roberts points out that the conditions for the compression of the available clouds cease to exist as the shock wave passes out of the gas. The spiral line of compression is quite narrow.

When the initially cool clouds of dust and gas collapse, they heat up. They should first become visible as murky clouds with a star deep inside, and these may be observable only in the infrared. Ultimately many of these young stars should become blue-white supergiants rich in the ultraviolet radiation capable of exciting the bright emission nebulas.

As the shock wave moves through the interstellar medium it will inevitably pass through some regions of below-average density where the compression of the shock wave is insufficient to produce conditions of collapse followed by the

OBJECT	AVERAGE RADIUS (LIGHT-YEARS)	ESTIMATED MASS (MASS OF SUN = 1)	ACCRETION OF MASS IN 100 MILLION YEARS (MASS OF SUN = 1)
LARGE CLOUD	12	2,000	1,000
LARGE GLOBULE	3	60	30
SMALL GLOBULE	.1	.2	.05

DARK NEBULAS of interstellar grains can be roughly grouped according to three categories: large cloud, large globule and small globule. Such objects are often associated with areas of strong infrared emission and are believed to be potential birthplaces of stars. Photographs on the next three pages are pictorial examples of objects in each of these categories.

COALSACK NEBULA in the southern Milky Way, photographed by the author with the Curtis-Schmidt telescope at the Cerro Tololo InterAmerican Observatory, is one example of a large cloud of interstellar grains. Although the nebula is visually very dark, it is actually quite transparent. Faint stars can be seen through the obscuring matter. Recent studies have shown that the most transparent sections of the Coalsack dim background stars by only one magnitude (two and a half times) in brightness; some regions, however, are much denser and dim stars by some five magnitudes (100 times). Many nebulas provide the right conditions for the formation of interstellar molecules such as formaldehyde, and may also eventually collapse into protostars or a cluster of protostars.

formation of protostars. It is therefore quite understandable that along a spiral arm we do not find a smooth distribution of young clusters and associations of stars. There should be voids between the lines of concentration of protostars and young objects. This picture of star formation seems to have very strong observational backing: the various stages are as plainly visible to the eye as plants in successive stages of development in a garden.

It would not be right to present the shock-wave theory as one that has received unquestioned acceptance. Some of the theoreticians working in the field are strongly opposed to the entire concept of the Lin-Shu density-wave theory of spiral formation. Their guess is that large-scale magnetic fields play a major role in the formation of spiral arms and protostars. All parties seem agreed, however, that spiral structure and protostar formation go together, whatever the mechanism that started it all may be.

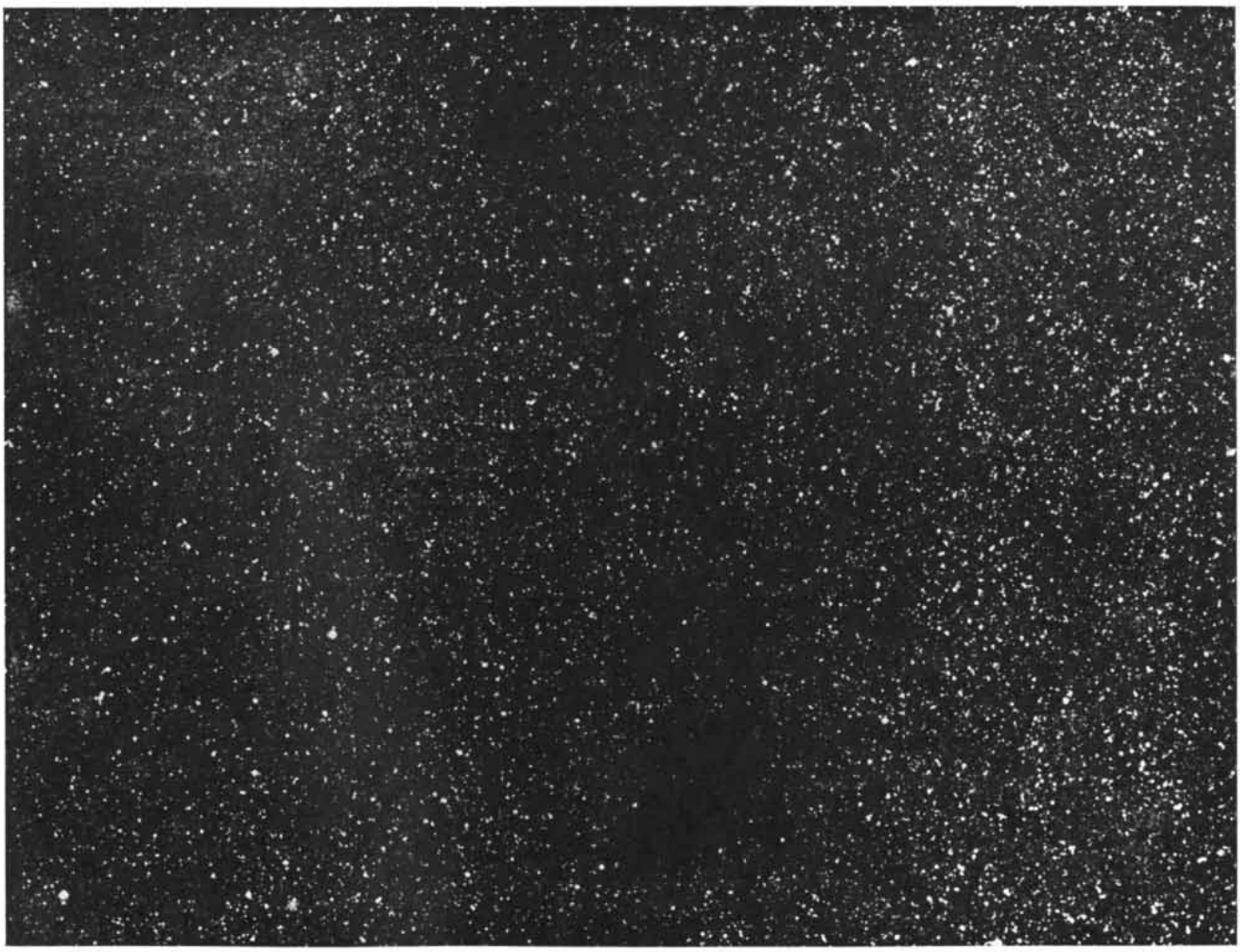
The dark nebulas are another class of objects that seem to suggest that star birth is taking place in the galaxy. They often appear as dark holes or clouds against a rich stellar background. Small interstellar grains are the chief known constituent of these nebulas. The reason the nebulas are dark is that the light from the stars beyond them is absorbed and scattered by the tiny grains. Interstellar gas is probably associated with the dust in such nebulas.

Dark Nebulas and Globules

The distance to a typical large cloud of interstellar grains can be estimated rather closely because the cloud reddens the light from stars behind it. It is therefore not difficult to distinguish between foreground and background stars, and the cloud can be assigned a rough distance. If the distance is known, the cloud's dimensions can be deduced from its apparent diameter in the sky.

A typical large cloud has a radius of some 12 light-years; the dust in it alone has a mass about 20 times the mass of the sun. Various studies have suggested that there is much gas associated with these objects; the amount of gas is generally estimated to exceed the mass of the cosmic dust by a factor of 50 or 100. In the course of time these clouds must sweep up a considerable amount of matter from the surrounding interstellar medium. It looks as though the amount of interstellar gas and dust swept up in an interval of 100 million years is equal to about half the estimated mass of the cloud. Most of the dark clouds should roughly double their mass in time intervals of the order of one cosmic year.

Globules of dust are objects of special interest. On photographs large globules often look like "holes in heaven." In a region with a rich and smooth background distribution of stars one suddenly encounters a darkened spot that looks like an area of low sensitivity in the photo-



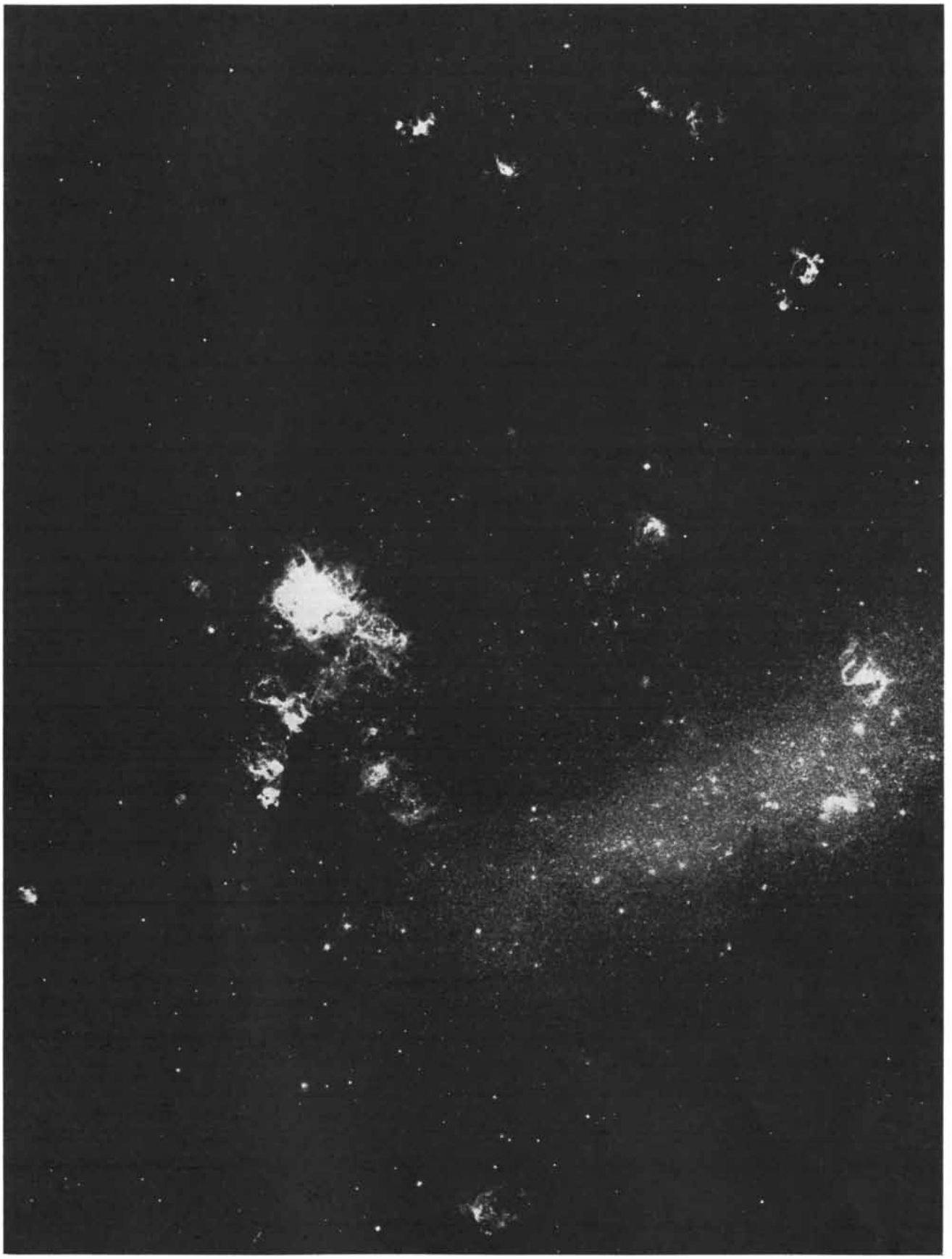
TWO LARGE GLOBULES appear as regions of low sensitivity of the film in this photograph in red light made with the 48-inch Schmidt telescope on Palomar Mountain. The globule near the top is No. 134 in the list compiled by Edward Emerson Barnard and the bottom one is No. 133. Given enough time, large globules will probably collapse into protostars and ultimately into actual stars.

graphic emulsion. There can be little doubt that these dark holes are roundish clouds or large globules of dust floating by themselves in interstellar space. Some long-exposure photographs with large modern telescopes show the background stars faintly coming through the obscuring matter; some globules must represent very dense small clouds of cosmic grains. Small globules are most often seen as tiny dark specks projected against the luminous background of a bright nebula. It is important to note that globules do not always accompany luminous nebulas, which suggests that conditions for their formation differ from one luminous nebula to another. No background light shines through the smallest dark specks. The masses assigned to globules are only guesses at a minimum figure. Moreover, it has not been possible to measure the gas content of globules. They contain such small amounts of interstellar gas that observational evidence for its presence is difficult to obtain. Some radio astronomers have discovered small, dense concentrations of the hydroxyl radical (OH), but these are mostly near emission nebulas and not at the position of the small globules.

Within 1,000 light-years of the sun, a small distance compared with the diameter of the galaxy, there are approximately a dozen large clouds and 100 fair-sized globules. It is not at all certain how many small globules there are. Small globules can only be seen projected against the bright emission nebulas, since they cover too small an area of the sky to be distinguishable against the normal background of stars. It is not known at present whether the small globules are selectively associated with the periphery of emission nebulas or are distributed more or less regularly in the galaxy. On the whole I favor the first suggestion. If small globules were present throughout the central plane of the galaxy, one would expect to see them projected against the luminous background of every bright nebula. As I have noted, that is not the case. The small globules are probably clouds literally rolled up into little dust balls by the pressure exerted by the expanding gas at the periphery of the nebula. The pressure of the ultraviolet radiation emitted by the hot Type O and Type B stars in the heart of each emission nebula probably assists in the formation of the globules. Dark clouds and globules seem to be units that have no choice except to gradually collapse into protostars or break up into clusters of protostars. Although the pressure waves emitted by a bright nebula may contribute to the formation of the smallest globules, the large clouds and large globules will probably collapse rather quietly on their own under the force of their own gravitation.

The Coalsack Nebula in the southern Milky Way is a fine example of a dark-nebula complex. It covers an area of the sky about five degrees square directly adjacent to the Southern Cross. Photo-

SMALL GLOBULES appear as tiny black spots near the top right corner of this photograph of the nebula IC 2294 made with the Curtis-Schmidt telescope at Cerro Tololo. Small globules are most often seen projected against the fringes of such luminous nebulas. Smallest of these globules have diameters of the order of the size of the solar system. They also are likely to collapse into protostars.



NEBULA 30 DORADUS (*left*) in the large Magellanic Cloud (*right*) is rich in bright nebulosity and young blue-white Type O and Type B stars, as shown in this photograph made by the author with the Curtis-Schmidt telescope at Cerro Tololo. This is one of the best examples of a region in which star birth should currently be taking place. The nebula is so large and luminous that if it were placed where the Great Nebula in Orion is now, it would fill the entire constellation of Orion and cast shadows on the earth at night.

graphs show it to be a region of below-average transparency. The stars that lie beyond the Coalsack shine through it with their brightness diminished on the average by from one to three magnitudes (2½ to 15 times). There are some very black spots in the nebula, generally oval in shape, through which no stars can be seen at all. We can almost imagine that the Coalsack Nebula is a dark cloud that is being fragmented into smaller units of cosmic grains; each of these units may eventually become a protostar. The Coalsack Nebula looks like the place where a star cluster is about to be born!

Supernova Explosions

A supernova explosion, in which a star nearly obliterates itself, is among the most spectacular of celestial phenomena. In A.D. 1054 Chinese astronomers observed a gigantic supernova explosion in the constellation Taurus at the position where we now find the Crab Nebula. There is a stellar cinder left behind, which is observed as a radio and optical pulsar and which appears to be a neutron star that has totally collapsed and is rotating on its axis in the incredibly short period of a thirtieth of a second. The Crab Nebula is 7,000 light-years from the sun and belongs to the Milky Way system. In 1971 attention was drawn to another supernova pulsar located about 45 degrees south of the celestial equator in the constellation Vela in the middle of the enormous luminous cloud recently named the Gum Nebula. The nebula is extremely diffuse and extends to a distance of 30 degrees or more from the central pulsar. The pulsar is some 1,400 light-years away, five times closer than the Crab Nebula [see "The Gum Nebula," by Stephen P. Maran; SCIENTIFIC AMERICAN, December, 1971].

The sun is less than 500 light-years from the nearer rim of the Gum Nebula (which is named for its discoverer, the late Australian astronomer Colin S. Gum). There are no known historical records to suggest that the explosion was observed in ancient times, but that is hardly surprising. The event must have taken place at least 11,000 years ago and possibly as much as 30,000 years ago. The explosion must have had two immediate effects on the surrounding interstellar medium. First, large amounts of gas enriched with elements heavier than helium must have been added to the interstellar medium around the supernova. Second, tremendous amounts of energy must have been transmitted to the surrounding interstellar medium in the form of explosive shock waves. The neutral atomic hydrogen that must have been there before the supernova outburst would have been ionized by the outpouring of ultraviolet energy. The supply of fresh energy is by no means exhausted; the tiny, rapidly rotating pulsar continues to pour it into the interstellar medium. Pulsars and supernovas are potentially a productive source of cosmic ray particles with energies high enough to help maintain the ionization of the surrounding medium. One of the most striking visible features of the Gum Nebula is its complex filamentary structure, the kind of effect one would expect from the passage of energetic shock waves through the interstellar medium.

It has been suggested that stars were formed or are still being formed in the region of the Gum Nebula. Several highly luminous and apparently young stars are close to the nebula and may have originated at the same time as the star that later became the supernova pulsar. Some of these stars are currently traveling at high velocity away from their apparent common point of origin.

The gaseous filaments visible in photographs of the Gum Nebula must represent highly condensed gas. Perhaps these filaments will ultimately break up into strings of protostars or young stars. Many workers in the field consider it likely that supernovas are conducive to the formation of protostars. The fact remains that no mass production of protostars has been observed in such regions. Moreover, there does not seem to be an abundance of young Type O and Type B stars near supernova remnants.

It may be naïve, however, to look for protostars or young stars in the regions close to recent supernovas. The collapse of a gas cloud into recognizable protostars or young stars is a process requiring at least 100,000 years, and in most cases as much as 10 million years. The Crab Nebula supernova explosion was observed less than 1,000 years ago; the supernova explosion at the heart of the Gum Nebula cannot have taken place much more than about 30,000 years ago. Even the youngest hot stars observed in the region near the Gum Nebula must have predated the supernova explosion by hundreds of thousands of years. The condensations from which these hot stars were born must have originated long before the recent supernova outburst that produced the pulsar. Supernovas may have much to do with the triggering of star birth, but so far we do not know precisely what it is.

Interstellar clouds have been studied in several quite different regions of the electromagnetic spectrum: X-ray, ultraviolet, visible, infrared and radio wavelengths. Such studies have yielded basic information on the properties of clouds, some of which may be on their way to becoming protostars. A variety of dark clouds and infrared objects have been discovered and classified. On the basis of such information we can ask: What kinds of mechanism are responsible for the development of a protostar and its collapse into a star?

The Formation of Stars

In a recent survey of theories of the formation of stars, Derek McNally of the University of London lists several processes that may be at work; in the end he favors star formation by collapse with gravity as the major cause. His conclusions are generally confirmed by the studies of other astrophysicists, notably C. C. Hayashi and his colleagues at Kyoto University. R. B. Larson of Yale University has drawn special attention to a process by which one of two things could happen. Either a cloud will collapse into many different units and form a cluster of stars or it will collapse much faster near the center than in its outer parts. The second alternative means that a star would be formed mostly from the material near the center of the cloud, and that a young star would be embedded in a large envelope of dust and gas. Such a star would have a truly murky atmosphere! Much attention is being given to how the extended atmosphere might collapse. It seems likely that the protostar would rotate and that this rotation would play an important part in holding it up for some time. In one way or another the protostar must get rid of some of the angular momentum that is stored in the rotating cloud of gas and dust. It can do so most readily by forming dusty shells around itself that in turn break up into planets. The theory for the formation of a protostar from a dust cloud seems to lead almost naturally to the formation of a planetary system.

Recent investigations of infrared objects strongly support the theory that stars are formed by collapsing clouds. Eric E. Becklin and G. Neugebauer of the California Institute of Technology have discovered an infrared point source near the heart of the Great Nebula in Orion that is almost surely a very young star. F. J. Low and D. E. Kleinmann of the University of Arizona have found a second object close to the same region; it seems to be a compact dust nebula, probably with a newborn star or cluster of stars near its center.

Evidence for nebulas enclosed in dust

shells has been forthcoming from radio-astronomical observations as well. Peter Mezger and his colleagues at the Max Planck Institute in Bonn have found a number of emission nebulas that emit strongly in radio wavelengths but are not detectable at visual wavelengths. The hypothesis is that these are "cocoon nebulas": brilliant nebulas embedded in clouds of interstellar grains. Their radiation in the radio region can pass through the surrounding dust clouds but their radiation in the visible region cannot. They might be observable in the infrared.

A class of intrinsically faint stars, the T Tauri stars, almost certainly represents a very early stage of stellar evolution. T Tauri stars vary irregularly in their energy output and show strong emission lines that are presumably produced in their extended outer atmosphere. The spectra of such stars also exhibit absorption lines that are formed deeper in the atmosphere. The lines are broad and fuzzy, indicating that the stars are either rotating rapidly or continuously ejecting mass. There is much evidence to support the hypothesis that gases are continuously flowing out of the atmosphere of a T Tauri star. Such stars are most often found in groups, generally near or within dark nebulas. The fact that they cluster together is so marked that V. A. Ambartsumian of the Byurakan Astrophysical Observatory in the U.S.S.R. gave them the name "T associations." E. Mendoza of the National Autonomous University of Mexico has found that T Tauri stars are strong emitters of infrared radiation.

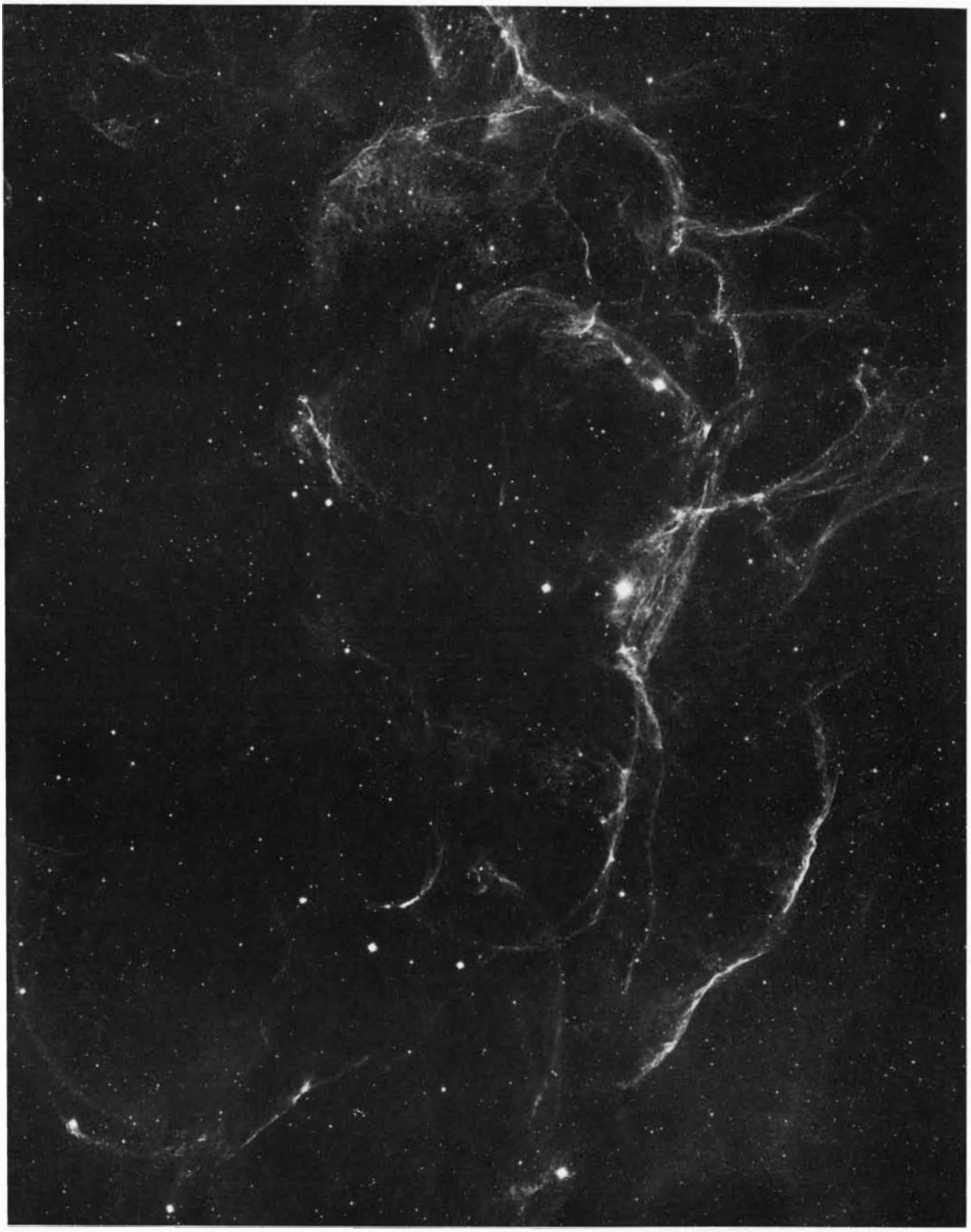
George Herbig of the Lick Observatory has apparently observed the formation of one truly new star, FU Orionis, which suddenly appeared in 1936. Quite recently Guillermo Haro of the Tonantzintla National Astrophysical Observatory in Mexico has drawn attention to a star that behaves much like FU Orionis: the faint variable star V 1057 in Cygnus. V 1057 Cygni has recently flared up and is now very bright in the infrared. Haro expresses the opinion that FU Orionis and V 1057 Cygni were originally T Tauri stars that have now advanced to the next evolutionary stage, which is represented by one or more characteristic long-term flare-ups. Herbig and Haro have discovered a number of small bright nebulas (now named Herbig-Haro objects) that are interspersed with the edges of dark nebulas, mostly in regions where T Tauri stars are abundant. T Tauri itself, the prototype for which the class is named, is embedded in such a nebula.

The general scheme I have described here is one favoring the formation of protostars through the process of gravitational collapse in clouds of interstellar gas and dust. I should point out that not every astronomer and astrophysicist favors this scheme. David Layzer of the Harvard College Observatory has developed the following theory of the related formation of stars and galaxies. In the beginning the mass of the universe was distributed quite irregularly. Fragmentation and clustering took place on a large scale, and there were some blobs of hot, dense plasma in which the gravitational field was much stronger than average and in which there were also large electric fields. Layzer believes conditions in these blobs would have been conducive to star formation. Another position has been taken by Ambartsumian and is strongly supported by Halton C. Arp of the Hale Observatories. They consider it likely that violent explosions in the nuclei of galaxies (including the nucleus of our own) may have much to do with the origin and maintenance of spiral structure. They suggest that associations of young stars along with interstellar gas and dust would be a direct result of the ejection of material by such explosions. So far neither the Ambartsumian-Arp theory nor Layzer's has been developed to the point where it can be checked in detail by observations.

It seems quite natural that star birth should be occurring now in the spiral arms of the Milky Way and of neighboring galaxies. Many dark nebulas and globules composed of interstellar gas and dust are seen almost in the act of collapsing into protostars or their close relatives. Objects that are either protostars or very young stars have also been observed. Infrared objects provide a natural link between small dark clouds and relatively normal stars; they may be cool, dense dust clouds with a star or a cluster of stars near the center. The cocoon nebulas may also supply newborn stars. The T Tauri stars seem to be the next stage and help to bridge the gap between the protostars and the young stars.

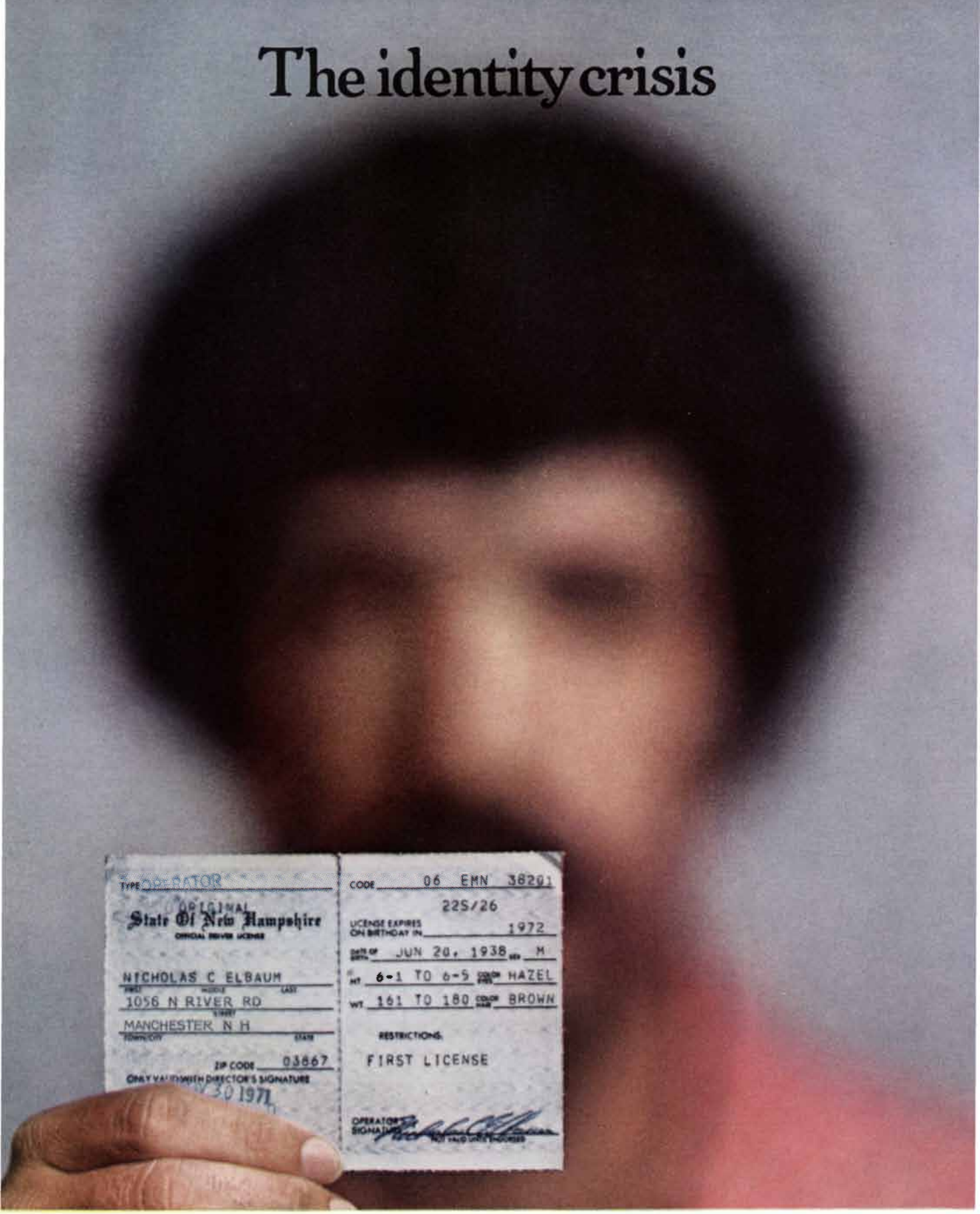
Comprehensive research is continuing on the problems of change in the galaxy and the related questions of the birth of stars and their early evolution. It is a good thing to describe the galaxy in all its majesty and to study the properties of its many components. The final aim, however, goes further. We want to know how the galaxy came into being, how the stars were formed and what the history and the future of the Milky Way system is.

GUM NEBULA, photographed by the author with the Curtis-Schmidt telescope at Cerro Tololo, has an overall diameter of



some 2,300 light-years. It is the remnant of a supernova explosion that may have occurred as long as 30,000 years ago. A pulsar, the stellar cinder of the supernova, is located near the center of the nebula in the constellation Vela. The delicate filaments must be gas that is highly condensed and enriched by metals from the supernova; they might ultimately break up into strings of young stars.

The identity crisis



Is the whole world "digit happy"? Think of your zip code. Area code. Driver's license. Charge cards. Bank cards.

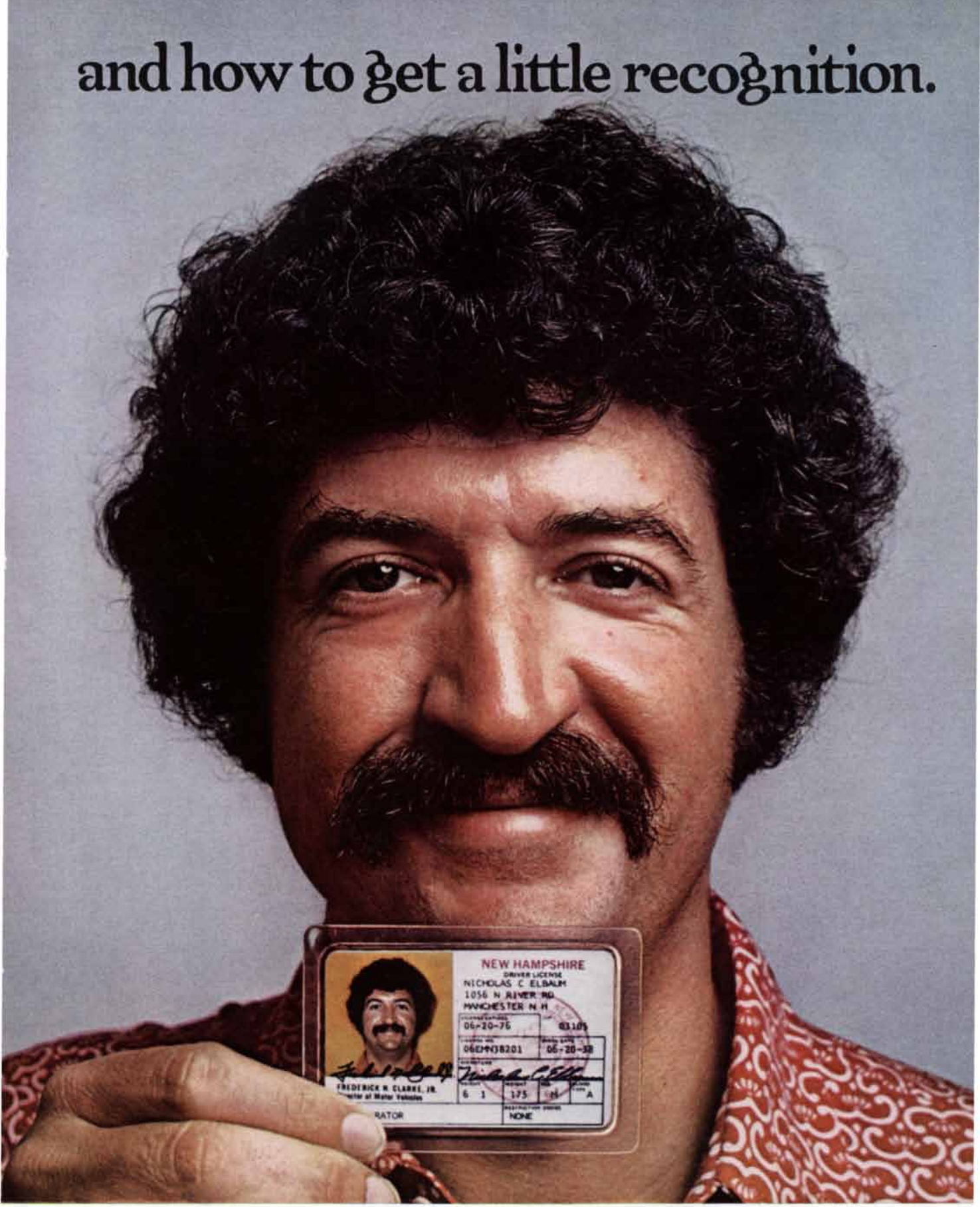
Does it ever make you feel faceless? Like just a set of numbers?

And when those cards get into the wrong hands, look what happens. Security problems. Thefts and frauds. Forged checks and phoney purchases. Even with all those numbers. Even with signatures.

But with Polaroid Portrait ID cards, you can be recognized for who you really are. On the spot. And you can have a feeling of security. Because they're virtually tamperproof. And no one else can use your face.

The result is that millions of people throughout the world carry them. And because the Polaroid ID-3 Identification System is portable and simple to operate, more and more organizations are using

and how to get a little recognition.



it. Whether businesses or banks, country clubs or colleges. For credit cards, security cards, driver's licenses or whatever. Like the license you see above. From New Hampshire, which is the latest state to use our system.

All of which means that more and more people will be showing their faces.

Why not see how the system works. For a demonstration, or just for more information, write to Polaroid Corporation, Dept. 57-244, Cambridge, Mass. 02139.

We'll show you how easy it is to help the individual get some recognition.

The Polaroid Portrait ID-3 System.

Polaroid®

DOCTOR-PATIENT COMMUNICATION

Observations in the pediatric clinic of a large hospital indicate that the physician often talks jargon or seems not to fully heed the patient's concerns. Mutual dissatisfaction is a frequent result

by Barbara M. Korsch and Vida Francis Negrete

The problem of dissatisfaction with the delivery of medical care in the U.S. is not solely a matter of inadequate financing or insufficient facilities and personnel. When all is said and done, it remains obvious that improvement of the funding and availability of medical service, although essential, would not in itself be a sufficient answer to the problem. The quality of medical care depends in the last analysis on the interaction of the patient and the doctor, and there is abundant evidence that in current practice this interaction all too often is disappointing to both parties. Systematic surveys confirm that there is widespread dissatisfaction among patients with doctors and among doctors with lack of cooperation by their patients.

Of the various factors that tend to contribute to this discontent, certainly one of the most important is poor communication between doctor and patient. In modern medical practice, which is now focused predominantly on technical knowledge, the physician may be engrossed in technical concerns and arcane terminology that mystify the patient. The traditional system of a close, long-term relationship with a "family doctor" is being replaced for the patient by short-term encounters with specialists. Moreover, many physicians no longer attach high importance to personal rapport with the patient; to some the "bedside manner" seems a concession to salesmanship not befitting a medical scientist. As we shall see, however, the failure to establish empathy with patients can be a serious bar to communication and patient response. (Indeed, the common neglect of this psychological factor by doctors may account in part for the flourishing of quacks and faith healers, whose main attraction for sick people is their skill in furnishing emotional reassurance.)

As an illustration of the critical importance of the personal interaction of patient and doctor, let us consider a not untypical case. A mother in a state of high anxiety about the persistent severe coughing of her infant son takes him to a hospital clinic. The physician, without greeting her or addressing her by name, asks a few matter-of-fact questions and examines the baby. He finds that the child has a postnasal drip of mucus (from an inflamed sinus) that is causing the cough. Without explaining the cause to the mother, the doctor simply prescribes nose drops, steam inhalation and perhaps an antibiotic and asks the mother to return with the baby in a couple of days for a follow-up examination. The mother, disappointed that the doctor has apparently shown no concern about her distress over the baby's cough, and unable to see how the prescriptions could stop or relieve the cough, buys cough medicine at the corner drugstore instead of using the prescribed treatment. She does not go back to the doctor for a follow-up. If the cough actually betokened a serious underlying condition, she has left the baby in jeopardy. The physician, for his part, writes her off as an uncooperative parent because she has not returned.

Over the past five years our research team at the Childrens Hospital of Los Angeles, associated with the University of Southern California School of Medicine, has been investigating the important problem of doctor-patient communication. Obviously this is a crucial but neglected aspect of medical care. Medical schools do an admirable job of teaching their students the complexities of medical science, but they still leave the learning of the "art" of medical practice to the individual's own initiative and intuition. This remains, in our view, a serious omission. However well informed a physician may be, and however conscientious about applying his knowledge, if he cannot get his message across to the patient, his competence is not going to be helpful. Moreover, more than half of a physician's working time in patient care, particularly in fields such as general practice, pediatrics and internal medicine, is spent on problems involving primarily psychological factors and the need for communication rather than technical knowledge. Consequently, detailed study of the process of communication between doctors (or other health professionals) and patients is essential for improvement of the delivery of medical care. By clarifying the process we can hope to develop principles that will make it possible to teach something about the "art of medicine" in medical school, instead of leaving it solely to intuition.

There have been many studies of factors influencing the response of patients to medical counsel and treatment: factors such as the patient's background, personality and previous medical contacts, and on the other hand the doctor's personal characteristics and training. For our study we had to find a setting and a technique that would enable us to examine selectively the doctor-patient communication process itself. How could we arrange to isolate this factor? We chose a simple but natural situation that seemed to answer our needs. The patient sample would be a large one, representing a mixed population of diverse ethnic, social, economic and cultural backgrounds. Similarly, the doctor sample visited by these patients would also be sizable. In each instance the patient

EMERGENCY CLINIC of the Childrens Hospital of Los Angeles was the setting for the observations of doctor-patient communication. In these photographs from a taped recording of a case the mother tells the physician that the child has digestive trouble (*1*) and then calms the infant (*2*) so that the doctor can proceed with his examination (*3*). After examining the baby the doctor advises the mother (*4*) to put the child on a diet consisting mainly of liquids for a day or two. The mother dresses the baby (*5*) and, just before leaving, listens (*6*) as the doctor emphasizes the importance of liquids over calories to prevent dehydration. Later the experimenters asked each mother for her reaction to the interview. They also checked to see if she followed the doctor's advice.

FATHER: How does his heart sound?

DOCTOR: Sounds pretty good. He's got a little murmur there. I'm not sure what it is. It's ...it uh...could just be a little hole in his heart.

MOTHER: Is that very dangerous when you have a hole in your heart?

DOCTOR: No, because I think it's the upper chamber, and if it's the upper chamber then it means nothing.

MOTHER: Oh.

DOCTOR: Otherwise they just grow up and they repair them.

MOTHER: What would cause the hole in his heart?

DOCTOR: H'm?

MOTHER: What was it that caused the hole in his heart?

DOCTOR: It's 'cause...uh...just developmental, when their uh...

MOTHER: M-h'm.

DOCTOR: There's a little membrane that comes down, and if it's the upper chamber there's a membrane that comes down, one from each direction. And sometimes they don't quite meet, and so there's either a hole at the top or a hole at the bottom and then... it's really...uh...almost never causes any trouble.

MOTHER: Oh.

DOCTOR: It's uh...one thing that they never get SBE from ...it's the only heart lesion in which they don't.

MOTHER: Uh-huh.

DOCTOR: And uh ...they grow up to be normal.

MOTHER: Oh, good.

DOCTOR: And uh...if anything happens, they can always catheterize them and make sure that's what it is, or do heart surgery.

MOTHER: Yeah.

DOCTOR: Really no problem with it. They almost never get into trouble so...

MOTHER: Do you think he might have developed the murmur being that my husband and I both have a murmur?

DOCTOR: No.

MOTHER: No. Oh, it's not hereditary, then?

DOCTOR: No.

MOTHER: Oh, I see.

(Someone whistling in the room)

DOCTOR: It is true that certain people... tendency to rheumatic fever, for instance.

MOTHER: M'mm.

DOCTOR: There is a tendency for the abnormal antigen-antibody reactions to be inherited, and therefore they can sometimes be more susceptible.

MOTHER: Oh, I see. That wouldn't mean anything if uh...I would...I'm Rh negative and he's positive. It wouldn't mean anything in that line, would it?

DOCTOR: Unh-unh.

MOTHER: No? Okay.

DOCTOR: No. The only thing you have to worry about is other babies.

MOTHER: M-h'm.

DOCTOR: Watch your Coombs and things.

MOTHER: Watch my what?

DOCTOR: Your titres...Coombs titres.

MOTHER: Oh, yeah.

PARTIAL TRANSCRIPT of an interview at the clinic shows a doctor failing to establish communication with a mother. First he raises the troubling specter of a hole in the child's heart and then does little to be sure that he has allayed the mother's concern. Without explanation he mentions SBE, which is subacute bacterial endocarditis. Later he advises the mother to "watch your Coombs" and evidently fails to understand that he has mystified her.

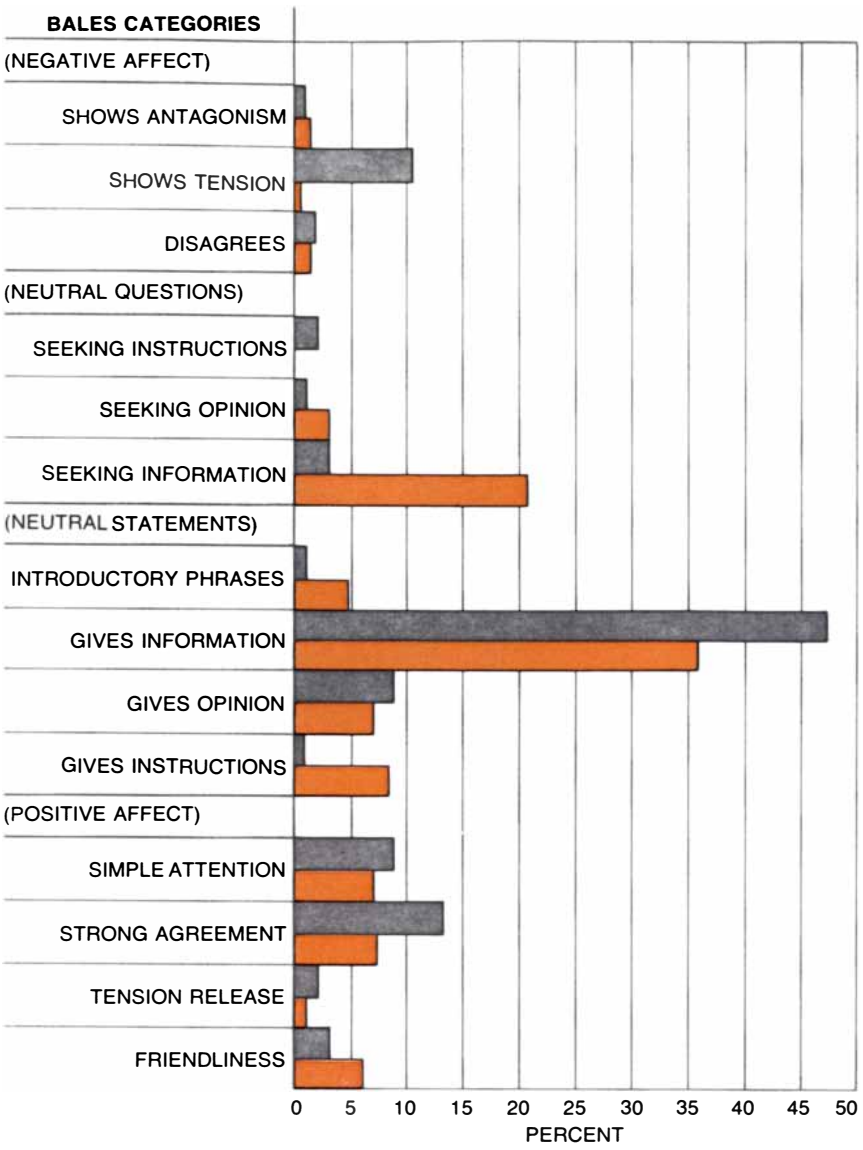
would visit a physician the patient had not met before and would present a case of acute illness (not previously diagnosed or treated) for which the physician could prescribe some definite treatment or course of action by the patient. The entire interview would be recorded, and afterward a member of our research team would follow up to learn how the patient responded to the interview and the doctor's instructions.

The situation we sought was well fulfilled by the emergency clinic of the Childrens Hospital. Children are brought to this walk-in clinic with a great variety of acute (but seldom catastrophic) illnesses or accidental injuries, invariably accompanied by a parent. The visit is usually short and generally yields a specific recommendation from the doctor to the parent. The clinic has a large staff of pediatricians, mostly young, well-trained, full-time residents. In our basic study we observed 800 visits by 800 different patients. Since the interaction was mainly between the child's mother and the doctor, we designated the mother as the "patient." Our standard procedure was to make an audio tape recording of the entire interview, then question the mother immediately afterward concerning what she had expected from the visit and what her reactions to it were and finally follow up later (within 14 days) to learn whether or not she had complied with the physician's instructions.

The setting and procedure provided a number of controls that minimized complication of the findings by extraneous factors, that is, variables other than the ones we wished to investigate. Since the patient came to a new physician to consider an acute illness not yet diagnosed or treated, the doctor-patient communication about the situation consisted only of the interchange between the two parties during this visit, uncomplicated by previous transactions or by any prior briefing of the parent about the illness. The large size of the samples (both of patients and of physicians) tended to correct for the bias of extraneous personal factors such as social or educational background when the responses of the group as a whole were considered. (Actually we found on analyzing the results that most patients, regardless of their personal background, responded to a given communication style in much the same way.) Another factor we had to consider was the possibility that the tape-recording of the visit might cause the doctor to depart from his usual style

and put his best foot forward, so to speak. It turned out, however, that in most cases the physician disregarded the presence of the tape recorder and behaved naturally. As a control, we omitted the use of the tape recorder in 300 of the 800 visits, and we found this apparently made no difference in the physicians' performance or the patients' reactions to the interview.

The findings from the 800 cases have since been supplemented with hundreds of other observations of doctor-patient communication, many of them involving routine checkups of well children, many not in clinics but in the private practice of pediatricians. In general these observations confirmed the validity of the conclusions from the basic study. It must still be borne in mind that the setting for that study was after all rather special: an emergency visit in a clinic on an acute but usually minor illness, generally with a young doctor of only brief pediatric experience (one to three years). The patient's response in a case of severe chronic illness or to a physician of long acquaintance might well have been very different. Our concern, however, was to look into the effects of particular modes of communication (or noncommunication) irrespective of other factors. The barriers to communication that were spotlighted in our study may occur in any setting, although they are not as common or as severe in private practice as they are in a single emergency visit to a clinic.

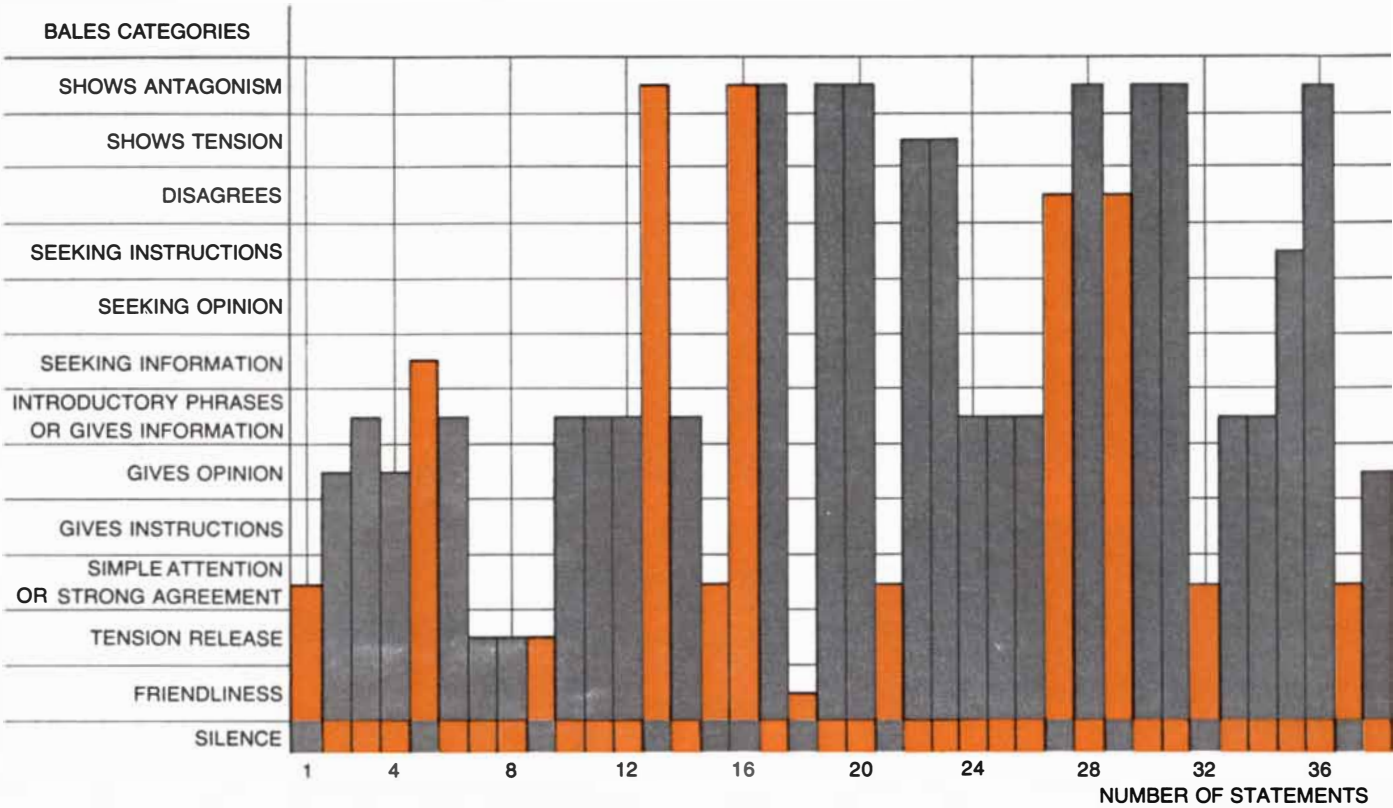


MEAN PROFILE of exchanges between physicians (*color*) and mothers (*gray*) is portrayed according to categories devised by the psychologist Robert F. Bales. For example, 2.9 percent of mothers' statements and 5.6 percent of doctors' were in the "friendliness" category.

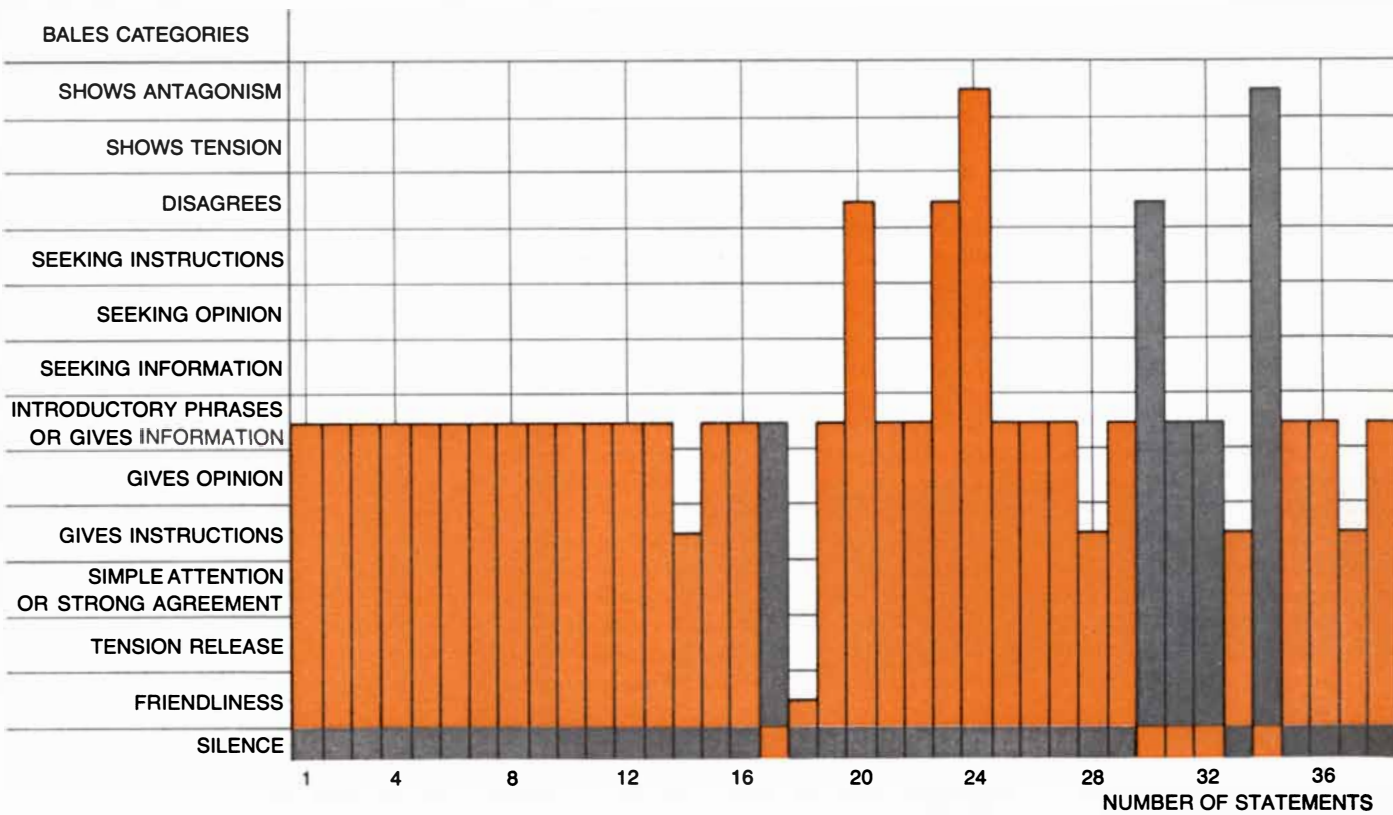
What, then, were the findings in detailed analysis of the 800 clinic visits? We consider first the mothers' evaluation of the conference with the physician. Immediately after the visit a member of our research team interviewed each parent to ascertain how she had felt about the child's illness, what she had expected of the doctor and how well satisfied she was with what he had said and done. Of the entire group, 40 percent were highly satisfied, 36 percent moderately satisfied, 11 percent moderately dissatisfied and 13 percent highly dissatisfied. That 76 percent of these anxious mothers were more or less satisfied with the doctor's performance in their brief encounter in the clinic is of course a reassuring finding. Their specific reactions, however, were less favorable. Nearly a fifth (149) of the 800 mothers felt they had not received a clear statement of what was wrong with their baby, and almost half of the entire group were still wondering when they left the physician what had caused their child's illness. The absence of an explanation of the cause is unnerving in such a situation, because the mother of a sick baby generally has a tendency to blame herself for the occurrence and needs specific reassurance.

The subsequent follow-up on how the mothers complied with the physician's instructions told a disquieting story. We took pains to obtain a true account by asking the mother searching but tactful questions (such as "When were you able to discontinue the treatment?") and by checking medicine bottles or the pharmacy when feasible. It turned out that 42 percent of the mothers had carried out all of the doctor's medical advice, 38 percent had complied only in part and 11 percent not at all. (In the remaining 8 percent of the cases the physician had not felt it necessary to give any prescription or advice.)

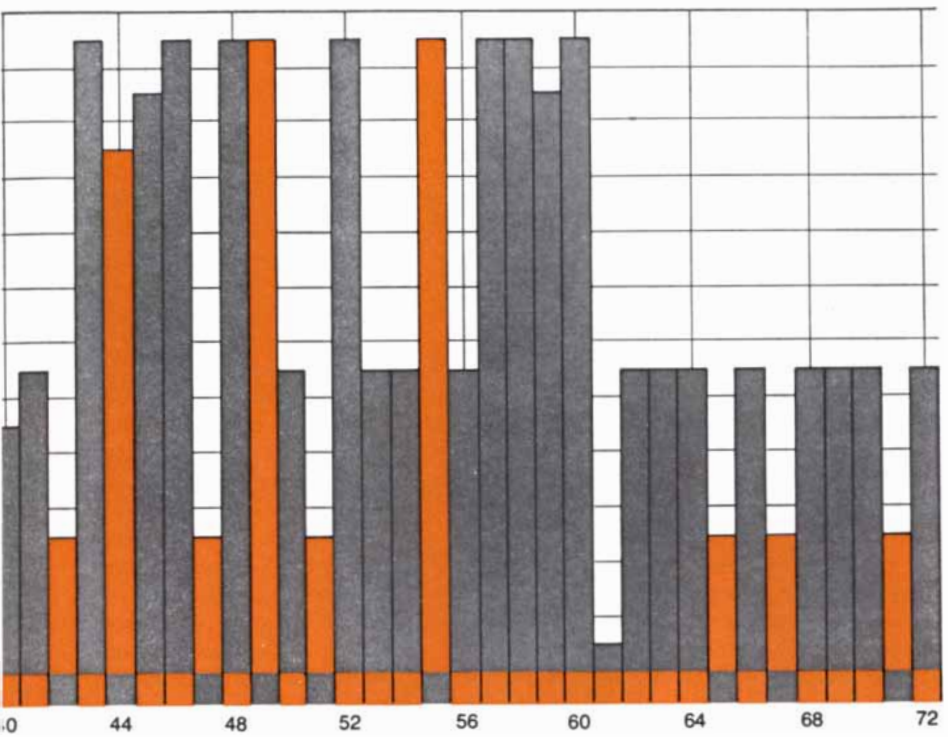
As was to be expected, we found a substantial correlation between the mothers' expressed satisfaction with the doctor's behavior in the visit and their compliance with his instructions. Of the highly satisfied mothers 53.4 percent cooperated completely with his advice, whereas only 16.7 percent of the highly dissatisfied patients did so. The fact that the correlation between compliance and satisfaction with the doctor was not consistently observed can be attributed to complicating factors such as the mother's view of the seriousness of the illness, the complexity of the physician's instructions, the difficulty of the prescribed



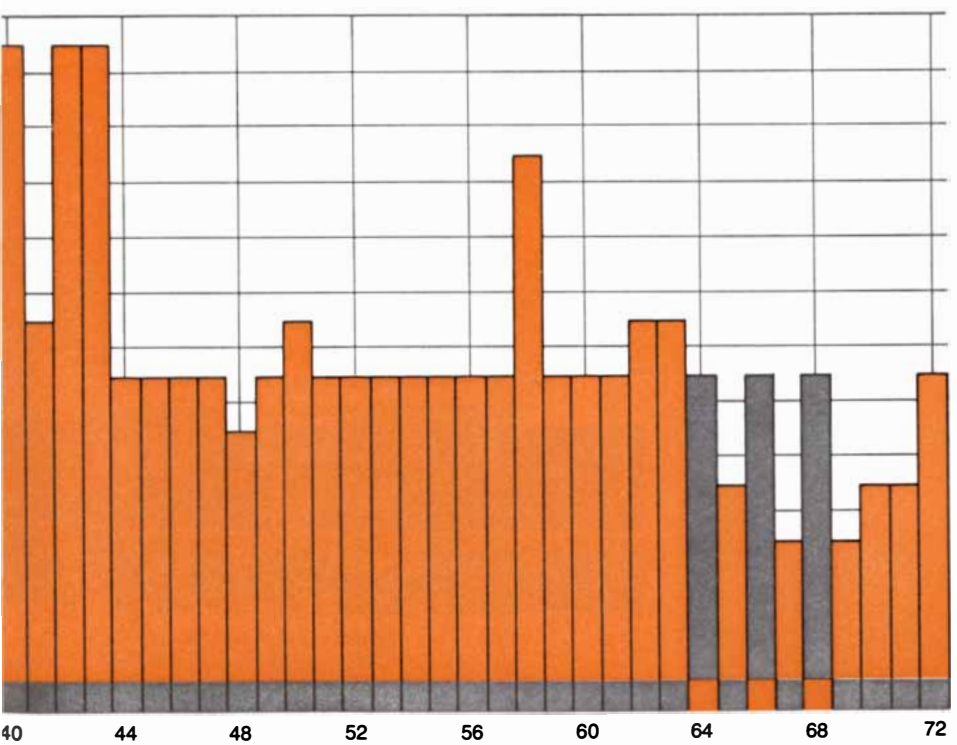
SEQUENCE OF INTERACTIONS between a physician (*color*) and a mother (*gray*) is traced, with the statements of each participant grouped according to Bales categories. In this segment of a longer interchange the mother made 50 statements, the doctor



SECOND SEQUENCE portrays a mother who appeared to be passive, remaining silent for long periods while the doctor talked. In a segment of 100 statements, most of which are shown here and grouped according to Bales categories as in the sequence depicted



22. The arrangement of Bales categories is such that the top three come under the heading of negative affect, bottom four, exclusive of "silence," are positive and others are neutral.



in the illustration at the top of these two pages, the doctor made 90, the mother 10. At only 14 times during the segment did the two have what could be regarded as an exchange.

treatment and various practical problems.

For light on the specific problems of communication between doctor and patient we now turn to detailed analysis of the content of their interchanges as recorded verbatim in the tapes. We coded the various features or elements characterizing their conversations and the patients' reactions and then submitted the data to analysis by computer. One of the tools we used was an adaptation of the "interaction process analysis" technique of the psychologist Robert F. Bales, which describes the content and tone of verbal interaction in terms of affect—positive and negative.

A question that immediately comes to mind with regard to a doctor-patient interview is the influence of the length of the session. It is commonly supposed that the more time the doctor can spend with the patient, the more satisfactory the results will be. No doubt part of the dissatisfaction with present medical care is attributable to the limited time harried physicians can give their patients. Surprisingly, however, the results of our study indicated that time was not necessarily of the essence. The 800 visits we examined varied in length from two minutes to 45 minutes, and we could find no significant correlation between the length of the session and (1) the patient's satisfaction or (2) the clarity of the diagnosis of the child's illness. Indeed, on examining some of the longest sessions we noted that the time was consumed largely by failures in communication: the doctor and patient were spending the time trying to get on the same wavelength!

The general impression that physicians tend to be too technical in language for their patients is strongly confirmed by our study. Terms such as nares, peristalsis and Coombs titre were Greek to the patients. A "lumbar puncture" was interpreted as meaning an operation to drain the lungs, and a reference to "incubation period" was taken to signify the length of time the sick child was to be kept in bed. A mother who was told that her child would be "admitted for a work-up" did not realize that he was to be hospitalized; when another mother was told by the physician that he would have to "explore," she had no idea he was talking about surgery. In more than half of the cases we recorded the physicians resorted to medical jargon. This did not necessarily leave the patient dissatisfied; some patients were impressed and even flattered by such language. It did, however, leave most of the mothers unenlightened about the nature of the

child's illness. One of the interesting findings was that satisfaction with the doctor's communication was not significantly higher among college-educated mothers than it was among those with less education.

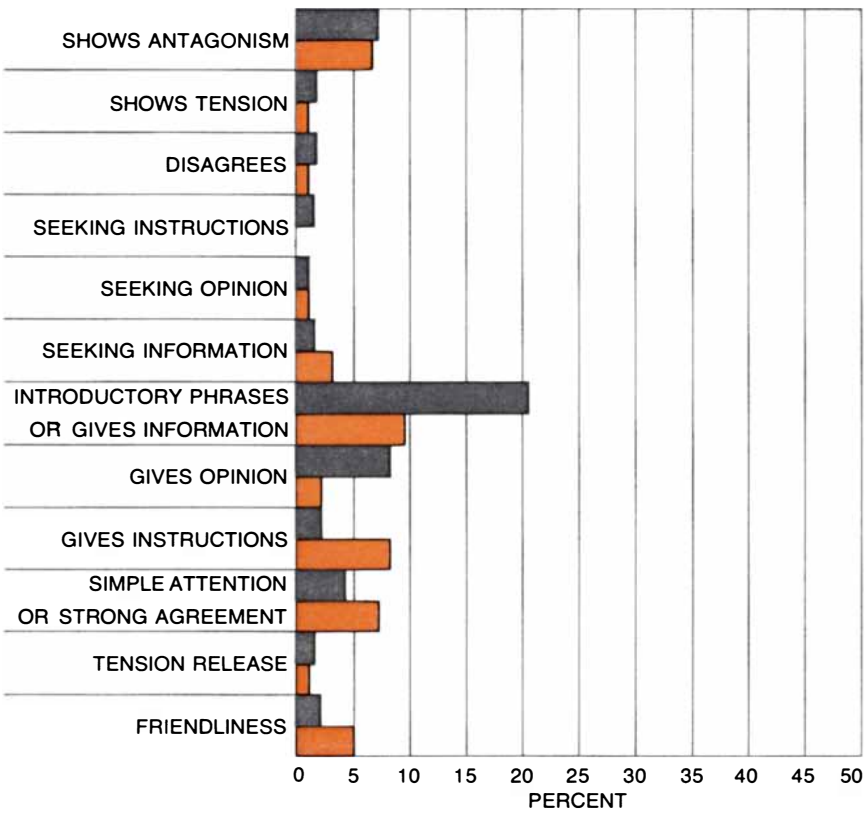
The language barrier was by no means the most serious bar to effective communication. The severest and most common complaint of the dissatisfied mothers was that the physician had shown too little interest in their great concern about their child. High among the expectations of mothers in coming to the clinic was that the doctor would be friendly and sympathetic not only to the child but also to the worried parent. The recordings show, however, that less than 5 percent of the physician's conversation was personal or friendly in nature. In most of the visits the physician gave no attention to the mother's own feelings and devoted himself solely to technical discussion of the child's condition. The disregard of the mother's concern must be considered an important hindrance to communication in the light of the fact that, as we found in the postvisit interviews, 300 of the 800 mothers held themselves in some way responsible for their child's illness. In a few instances the physician even expressly blamed, or appeared to blame, the mother. In one case a physician remarked to the child, perhaps in jest, "Stevie, it's your mother's fault that you have this high fever." The mother later voiced great distress over this to our interviewer.

A frequent cause of dismay for the mother was the physician's total disregard of her account of what chiefly worried her about the child's illness. When, for instance, a mother repeatedly tried to interest the doctor in the fact that her child had been vomiting, he ignored her remarks and persisted in asking her about other symptoms, which, as she did not realize, related to the same basic problem—dehydration of the child. Another mother feared that her child's febrile convulsions might cause permanent damage to the brain, but she did not succeed in engaging the doctor's attention to this concern. Among the 800 mothers, 26 percent told interviewers after the session with the doctor that they had not mentioned their greatest concern to the physician because they did not have an opportunity or were not encouraged to do so.

Under such circumstances there was frequently a complete breakdown of communication. Some patients were so preoccupied with their dominant concerns that they were unable to listen to the physician. Some even reported that the physician had failed to examine the child adequately or to give a prescription, although the tape-recorded account of the visit attests that he did in fact do so. Among mothers who felt that the physician had not understood their concerns 68 percent were dissatisfied with the visit, whereas of the 625 mothers who reported the physician *had* understood, 83 percent were satisfied.

We have mentioned that 149 patients reported they had not been told what was the matter with their sick child. The recordings of the visits show that in many cases the physician did indeed fail to provide a clear diagnostic statement, and often he offered no prognosis. Diagnosis was of course one of the main expectations that had brought each mother to the clinic. Many of the mothers complied with the doctor's medical advice even when no diagnosis was given. Understandably, however, omission of such important information did not tend to inspire confidence in his prescriptions. Only 54 of the 800 patients seriously questioned the physician's technical competence (in their postvisit interviews with us), but failure to show a friendly interest or to fulfill their other expectations was a significant deterrent to compliance with his instructions. Of the patients who felt that the physician had not met any of their expectations, 56 percent were grossly noncompliant.

On the positive side, the recordings of the hundreds of doctor-patient conversations clearly identified specific forms of discourse that made for good communication and patient satisfaction. One of these, of course, was expression by the doctor of friendly interest in the "patient" with whom he was conducting the conversation (that is, the mother). Most of the physicians believed they had been friendly, but fewer than half of the patients had this impression, and 193 of them reported that the doctor had been strictly businesslike. Attention to the mother's worried concerns had a high correlation with success in satisfying her and obtaining her compliance with advice. This suggests that a physician can quickly establish fruitful communication with the patient by opening the conversation with questions such as: "Why did you bring Johnny to the clinic? ... What worried you the most about him? ... Why did that worry you?" A brief but friendly discussion of the patient's con-



ENTIRE INTERVIEW between the doctor (*color*) and the mother (*gray*) involved in the case charted at the top of the preceding two pages is summarized according to Bales categories. In the interchange the number of statements by mother was 244 and by doctor 198.

cerns, however irrelevant or irrational they may seem, can perform wonders in reassuring her and winning her cooperation. Even when the physician was not able to fulfill all the mother's expectations, a demonstration of warm concern and individualization of his advice achieved satisfying results. The patients reacted poorly to impersonal or institutional expressions such as "We don't hospitalize children with impetigo" or "We keep most cases of pneumonia under observation in the clinic." On the other hand, patient rapport and cooperation thrived on specific instructions, expressions of trust in the mother's caretaking ability and offers of continued interest such as "Call me anytime" or "We'll check Johnny again tomorrow."

Detailed study of the recordings with the aid of Bales's method of interaction process analysis brought forth a number of significant findings, some of them unexpected. The verbatim records showed that on the average the doctor did more talking than the mother, which proved to be a surprise to these physicians and probably comes as news to the medical fraternity generally. The session tended to have a more successful outcome when the patient had an active interchange with the doctor than it did when she remained passive and asked few questions. In general the patients were disappointingly reticent about asking questions or opening up lines of inquiry, in view of the anxiety and desire for more information that they expressed to interviewers afterward.

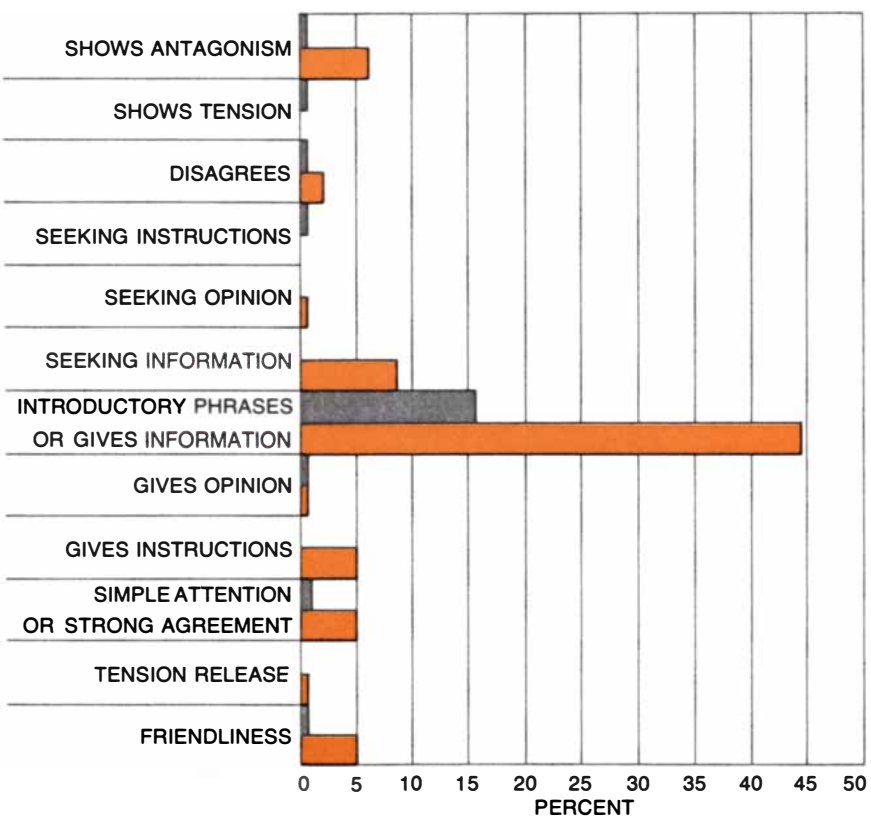
It may be significant in this connection that in some recordings the doctor-patient conversation comes to a distinct breaking point, after which no real communication takes place and one or the other participant is reduced to mechanical uh-huhs or yeses. In other cases the physician is found to fall into repeating statements several times and showing increasing impatience and irritation. These two types of situation probably reflect great tension on the part of the patient as well as the collapse of communication.

The verbal records give relatively few obvious signs of affect. Civilities between the parties, such as introducing themselves or addressing each other by name, are uncommon. The interchanges consist mainly of neutral, informational statements [*see illustrations on pages 70 and 71*]. Nevertheless, the tone and emotional content of the encounters is amply evidenced. One noteworthy finding is that, whereas less than 6 percent of the doctor's communication to the

mother carries positive affect (in the form of friendly remarks, joking, agreement, support), 46 percent of his conversation with the child is of this nature. Plainly the physician not only identifies with the child rather than with the mother as the patient but also feels a need to give more reassurance to the child than he would to an adult. The results in this study, however, indicated that the physician's friendliness to the child had only a slight influence in heightening the mother's satisfaction with the visit or getting her to follow his medical advice. It was his attitude toward her that counted most.

If the physicians rarely show positive affect to the mother in these records, by the same token they seldom show negative affect in the form of disapproval, criticism or hostility. When the doctor does express negative feelings, the mother is likely to be dissatisfied with the visit and fail to comply with his advice. Conversely, a substantial showing of positive affect by the doctor to the mother enhances her satisfaction and compliance. This finding has a bearing on a controversial issue in medical practice. There is a widely held belief that the doctor should maintain a certain social distance from the patient to strengthen his image as a figure of authority, and some physicians go so far as to use scare techniques to obtain compliance with their advice, threatening dire consequences if it is not followed. Some patients seek out this type of doctor; they may in fact need such treatment. In our investigation, however, we came across few such individuals. Friendly treatment of the patient generally had favorable results; harsh treatment tended to yield poor results. And there was a direct statistical relation between the amount of nonmedical (that is, sociable) conversation between doctor and patient and the patient's satisfaction with the encounter with the doctor.

The patients exhibited considerably more negative affect than the physicians did. Very few of the mothers openly expressed hostility or resentment to the doctor; expressions of negative feelings usually took the form of statements indicating nervousness or tension. Such statements ran as high as 45 percent of all the utterances by the mother in some cases, and in the total sample they amounted to 10 percent of all the patients' statements. In a large number of these instances the physician did not offer any reassuring response to the mother's indication of anxiety.

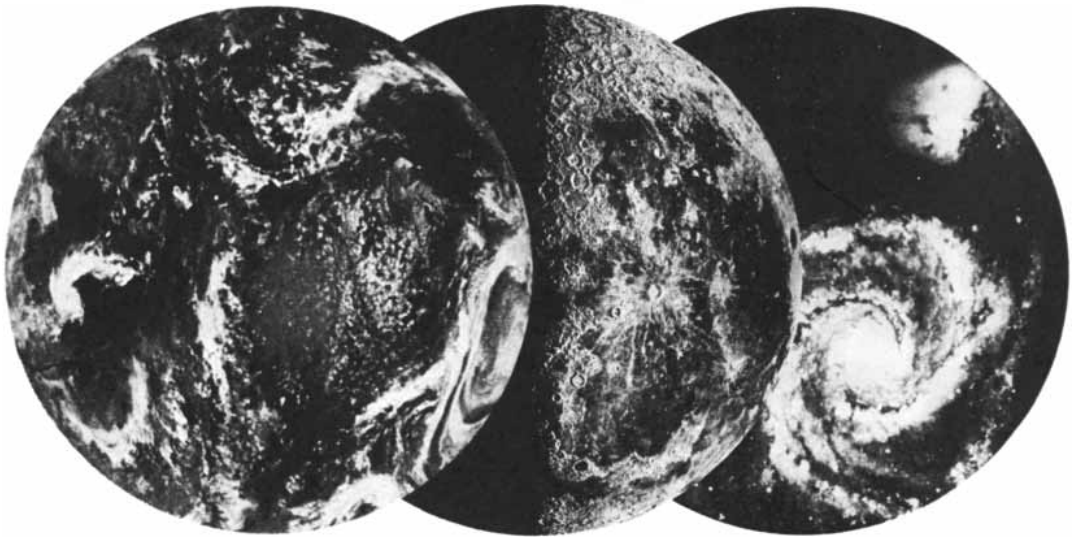


NATURE OF REMARKS by doctor (*color*) and mother (*gray*) involved in the case charted at the bottom of pages 70 and 71 is summarized. Doctor made 184 statements, mother 46.

Our exploration of the communication aspect of health care in some depth has opened up encouraging prospects for relatively simple ways to improve the delivery of this care. It is certainly not our intention to undermine public confidence in the medical profession; on the contrary, the lessons learned from analysis of the communication problem can go far to help the profession gain support and strengthen its performance. Whereas other problems that stand in the way of delivering health care satisfactorily to the entire population seem to call for reorganization of the entire social structure and basic personality changes in the people, the communication problem can be solved more easily.

The shortcomings in communication that we have examined in the clinic situation after all reflect a pattern that is common in medical practice generally. Furthermore, the need for understanding the problem of communication and coping with it is increasing as the delivery of medical care is taken over more and more by specialized professionals and technicians, so that the patient must relate to a galaxy of different health workers. Unquestionably attention to effective communication, a skill that should not be too difficult for any trained person to master, could make a valuable contribution to the quality of health care and its availability to the general population.

With the technique of detailed analysis that our research team has used in examining verbal communication we are looking into certain other aspects of medical practice. We have begun to make video tapes of medical visits, which enable us to study nonverbal communication and to document the "instrumental" (as distinguished from "expressive") performance of doctors, including examinations of the patient. When a body of data on all these matters, expressive and instrumental, has been developed and units of behavior in the process of health care have been clearly defined, there will be a more solid basis for establishing optimal standards and comparing actual performance with these standards. It will then be possible to measure the quality of health care, to relate the elements of the process to results in patient health and to evaluate the contribution of the social and emotional aspects of patient care. These aspects may well be found to have a far weightier influence in preserving health and well-being than they are credited with now.



Flexible Contemporary Pertinent

SCIENTIFIC AMERICAN Offprints

More than 800 articles are now available in the SCIENTIFIC AMERICAN Offprint series. New titles are being added constantly to this important collection of reprints from SCIENTIFIC AMERICAN. Each Offprint costs 25¢. This sampling of titles will show you how much pertinence and educational flexibility a quarter can buy.

125. Brazier: The Analysis of Brain Waves
144. Cole: The Ecosphere
1023. Kolff: An Artificial Heart Inside the Body
1068. Pirie: Orthodox and Unorthodox Methods of Meeting World Food Needs
1129. Tietze & Lewit: Abortion
1170. Berns & Rounds: Cell Surgery by Laser
1197. Singer: Human Energy Production as a Process in the Biosphere
1199. Bodmer & Cavalli-Sforza: Intelligence and Race
309. Gardner: Can Time Go Backward?
311. Alfven: Antimatter and Cosmology
319. Weisner & York: National Security and the Nuclear-Test Ban
480. Sachar: Behavioral Science and Criminal Law
483. Barron, Jarvik & Bunnel: The Hallucinogenic Drugs
498. Levine: Sex Differences in the Brain
506. Calhoun: Population Density and Social Pathology
524. Grinspoon: Marihuana
611. Leontief & Hoffenberg: The Economic Effects of Disarmament
618. Haagen-Smit: The Control of Air Pollution
626. Taeuber: Residential Segregation
631. Lewis: The Culture of Poverty
633. Comer: The Social Power of the Negro
650. Alexander: The Cost of World Armaments
656. Converse & Schuman: "Silent Majorities" and the Vietnam War
885. Wenk: The Physical Resources of the Ocean
889. Hess, Kovach, Gast & Simmons: The Exploration of the Moon

Students' sets will be collated at no additional cost. Upon request, instructors will receive a complimentary copy of each Offprint ordered in quantity for class use.

For a FREE descriptive catalogue, write to

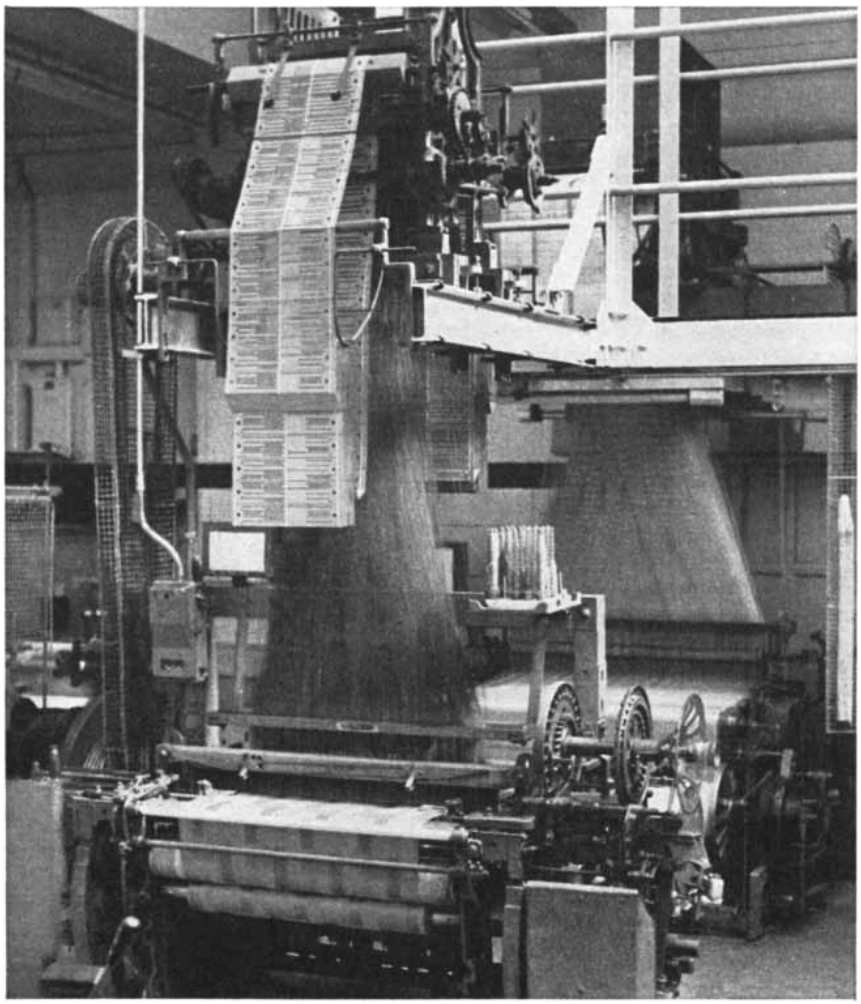
W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104
58 Kings Road, Reading RG1 3AA, Berkshire, England

ORIGINS OF THE BINARY CODE

This key element of modern computing systems can be traced as far back as the 17th century, when Francis Bacon devised a binary scheme of lettering for encoding his secret messages

by F. G. Heath

The binary system of numbers has become an important element in modern technology, being employed as a means of calculation in digital computers and as a means of controlling a considerable variety of machine tools. It seems safe to say, however, that few of the people who are familiar with the system could trace its history or identify the men who made the important discoveries in the field. Here I shall describe four such discoveries, the contributions of four men who approached the subject from different directions. The men were Francis Bacon (1561–1626), whose *omnia per omnia* code for conveying secret messages in apparently innocent communications was binary; Joseph Marie Jacquard (1752–1834), who designed a system of binary-coded punched cards for operating looms; George Boole (1815–1864), the English mathematician whose algebra of propositional calculus forms the basis of the modern design of computer logic, and Emile Baudot (1845–1903), a French engineer whose cyclic-permuted code (now often called the Gray code in the U.S.) represented a major advance in telegraphy.

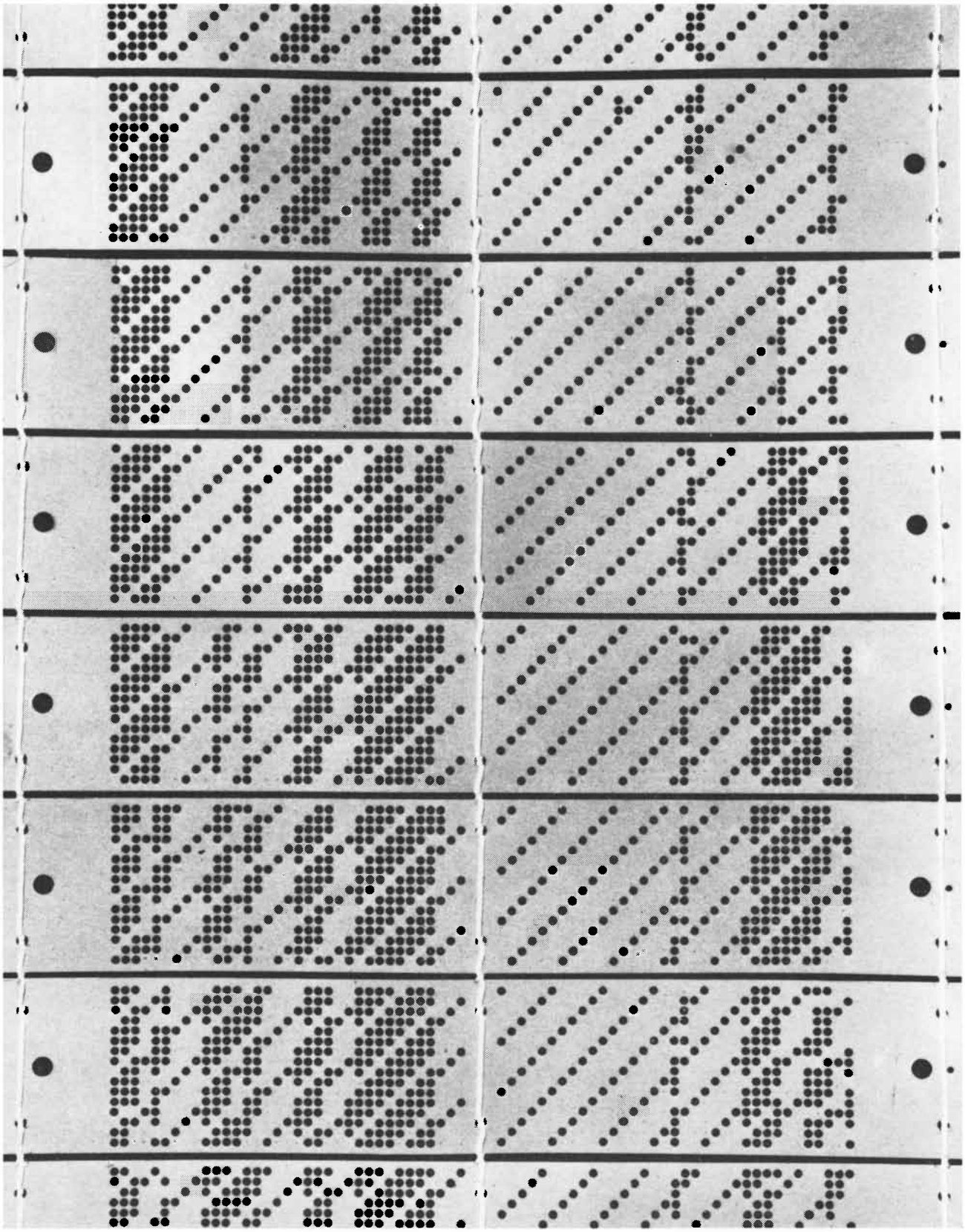


JACQUARD LOOM weaves by means of punched cards that control the lifting of warp threads as weft threads pass through. Such looms, developed in France early in the 19th century by Joseph Marie Jacquard, were the first binary-controlled production machines.

The binary (two-part) system of numbers takes its name from the fact that it employs only two symbols: 0 and 1. Units of the more familiar decimal (10-part) system, which is based on the numbers 0 through 9, are expressed in the binary system as follows:

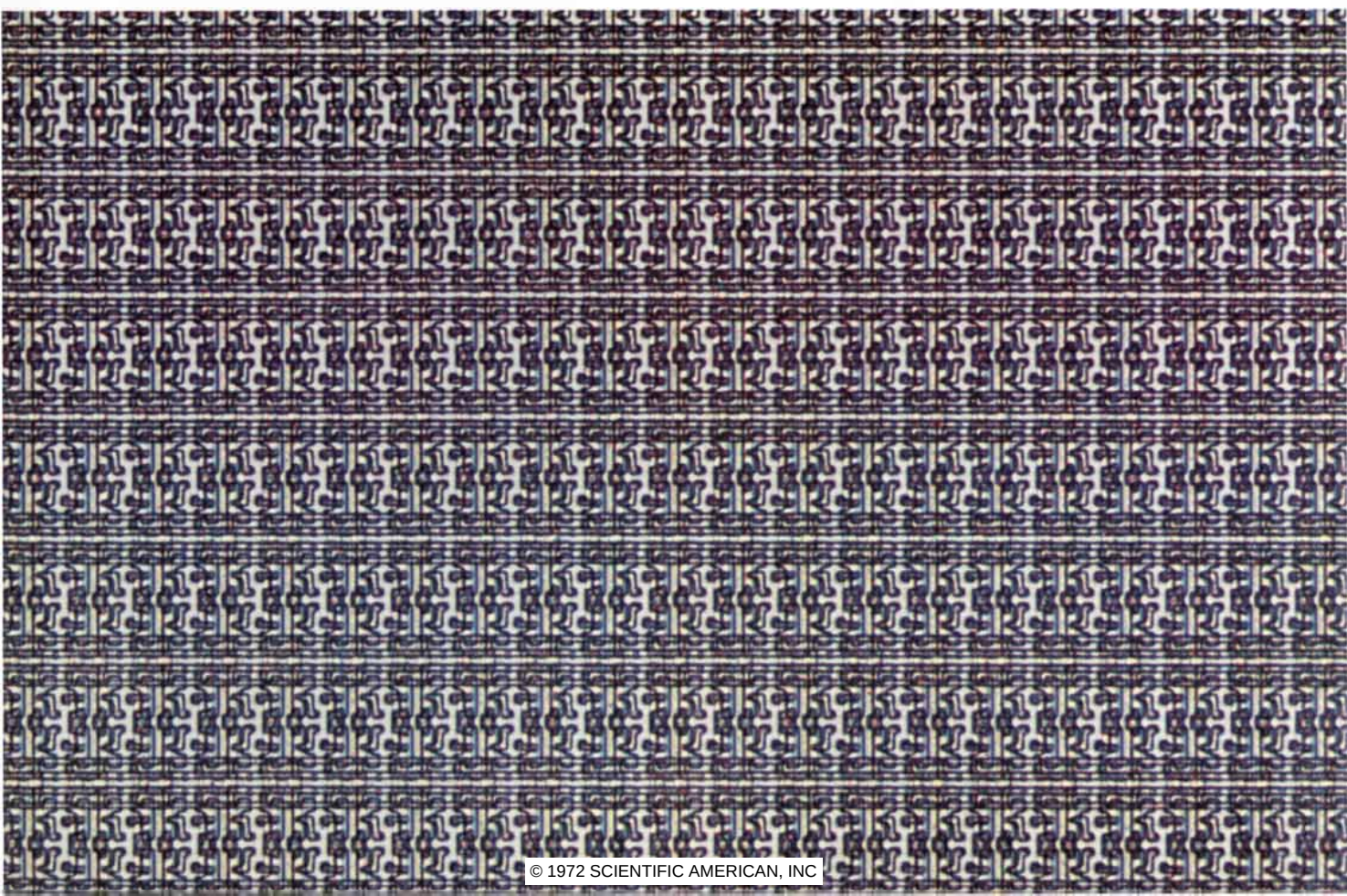
Decimal	Binary
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001

The reader will see that each progression in the system is to the next-largest number that can be made from 0 and 1. In converting a somewhat larger number, such as 41, from decimal to binary the same progression produces 101001, which can be converted back into decimal by reading the binary numbers from right to left as increasing powers of 2. Therefore in this particular example one ascertains that there is one 1, no 2, no 4,



PUNCHED CARDS control the operation of a Jacquard loom at Burlington House in New York. The function of the cards is to control the lifting of warp, or lengthwise, threads so that the shuttle carrying weft, or crosswise, threads can pass through the warp in such a way as to achieve a given pattern in a piece of woven material. The cards pass over a quadrangular block of wood, each face of which has a perforation opposite each of a number of horizontal needles associated with hooks that raise the warp threads. Wherever there is a hole in a particular card the associated needle will enter the corresponding hole in the block, and the warp thread controlled by the needle will be lifted. Where there is no hole the needle is blocked, and the corresponding hook is tilted aside so that its warp thread is not lifted. The number of cards required to weave a pattern equals the number of weft threads in the pattern.

FABRIC AND CIRCUIT PATTERNS have a resemblance resulting from their connection with binary systems. Above is an enlargement of a fabric woven on a Jacquard loom; below, an enlargement of a 1,024-bit memory circuit made by Fairchild Semiconductor. In a fabric the binary relation at a given point is whether a horizontal thread or a vertical thread is on top. The binary relation in the integrated circuit is whether a given area of an element is conducting or not, that is, whether it is metallized or not.



one 8, no 16 and one 32, or $1 + 8 + 32 = 41$.

Binary coding has had only a minor role in the ordinary transactions of mankind, plainly because the long trains of 0's and 1's have a hypnotic effect and are hard to recognize and copy. Probably the only use of the binary system in unmechanized computation was the "peasant multiplication" scheme, in which the two numbers to be multiplied are set down (as in 10×12) and the number on the left is halved (ignoring fractions) while the number on the right is doubled; the problem-solver then crosses out in the right-hand column all numbers that are opposite even numbers in the left-hand column and concludes by adding the numbers that remain on the right. The example of 10×12 would proceed as follows:

10	~~12~~
5	24
~~2~~	~~48~~
1	96
	120

This method employs binary processes but operates on decimal notation.

The great value of the binary system is in operations that are binary by nature: on or off, open or closed, true or false, go or no-go and so on. A given electronic component of a computer is either on or off—each state being described as a "bit" of information in the handy word derived by truncating the phrase "binary digit." That is why the binary system lends itself so effectively to the rapid calculating done by computers.

Bacon's contribution was not in computing but in coding. Nonetheless, he was the first person on record who consciously devised a binary code and described its properties. His purpose was to develop a cipher for his secret diplomatic messages. First he established for each letter of the 24-letter English alphabet of his time (*v* and *j* were not then in the alphabet) a five-letter binary code made up of various combinations of *a* and *b:* the letter *a* was *aaaaa, b* was *aaaab* and so on through the alphabet without repeating any five-letter grouping [*see illustration on next page*]. (He could equally well have made 0 and 1 the symbols of the code.)

Bacon's second feature was what he called a "biformed alphabet," wherein two slightly different kinds of type were specified for printing the innocent message. One set of type represented the *a* of his code, the other set the *b*. For example, one capital *G* would be drawn from a certain font of type and would represent an *a;* the second capital *G* would be drawn from another font and would represent a *b*, and two lowercase *g*'s would be similarly differentiated.

Supplying an example, Bacon proposed encoding the message *Fuge,* which is Latin for "Flee." From the five-letter groupings one ascertains that *f* is *aabab, u* is *baabb, g* is *aabba* and *e* is *aabaa.* Now one must formulate an exterior message that encompasses the interior message; Bacon's exterior message was *Manere te volo, doneo venero.* The *M* and the *a* would be printed in type from the font representing *a* in the code, the *n* from the *b* font, the *e* from the *a* font and the *r* from the *b* font, thereby yielding to the possessor of the code the letter *f*. Proceeding through the exterior message, one would emerge (assuming that the typesetter always drew from the proper font) with the following substitution for the message: *aababb aa bbaa bbaaa baa[aaa].* Dividing the letters into groups of five, one would obtain the coded version of *Fuge.*

Bacon mentioned the code briefly in 1605 in *Of the Advancement of Learning.* He described it fully in 1623 in *De augmentis scientiarum.* The central aim and binary nature of the code were set forth as follows in a passage of the latter work (as translated in 1640 by Gilbert Wats): "Neither is it a ſmall matter theſe *Cypher-Characters* have, and may performe: For by this *Art* a way is opened, whereby a man may expreſſe and ſignifie the intentions of his minde, at any diſtance of place, by objects which may be preſented to the eye, and accommodated to the eare; provided thoſe objects be capable of a twofold difference onely; as by Bells, by Trumpets, by Lights and Torches, by the report of Muskets, and any inſtruments of like nature."

In Bacon's opinion the binary arrangement provided the most effective cipher. He resorted occasionally to others, however, including a wheel cipher and a key cipher, which he seems to have thought of himself. Such a cipher employs key words in a message to bring in paragraphs from unrelated places, the whole assembling into a secret message. The scheme now appears in large computers under the name of the descriptor or code-word indirect-addressing method.

In a similar way Jacquard's loom is the ancestor of the many digitally controlled machine tools employed in modern technology. A more recent ancestor is the player piano. Neither of these devices was a true prototype of the modern equipment, since their processes are binary by nature whereas the modern machines are often involved in far more complex processes. In a loom a particular warp thread is lifted or not lifted when the shuttle goes through; in a piano a given note is sounded or not sounded at any particular moment. Nonetheless, the Jacquard loom was the first binary-controlled production machine to be widely used.

The somewhat contradictory accounts of the loom's development can be summarized as follows. When patterns began to be important in woven fabrics, the textile industry first resorted to a human drawboy to lift the appropriate threads of the warp before the weft passed through. (The warp threads in a woven fabric run lengthwise, the weft or woof threads crosswise.) In both Italy and France mechanical drawboys were developed during the 17th century; these devices reduced the labor, but the loom still required an operator. In 1725 Basile Bouchon devised a scheme whereby perforated paper with a set of needles controlled the warp threads. In 1728 Jacques de Falcon advanced the idea of a loom that had one card for each shedding (warp-dividing) movement, and in 1745 Jacques de Vaucanson made a working loom that included the ideas of Bouchon and De Falcon.

By 1801 this loom was in disrepair, and Jacquard, who apparently had established a reputation as an expert by having some loom equipment in an exhibition, was summoned to Paris by Napoleon and asked to go to work on the De Vaucanson machine. His contribution was to replace the perforated cylinder that carried the punched cards with a prism arrangement that allowed for many more cards being presented in sequence at the appropriate place in the loom. This key modification was backed up by excellent engineering, which was important because the whole idea depends on close mechanical tolerances—a novel concept in those days.

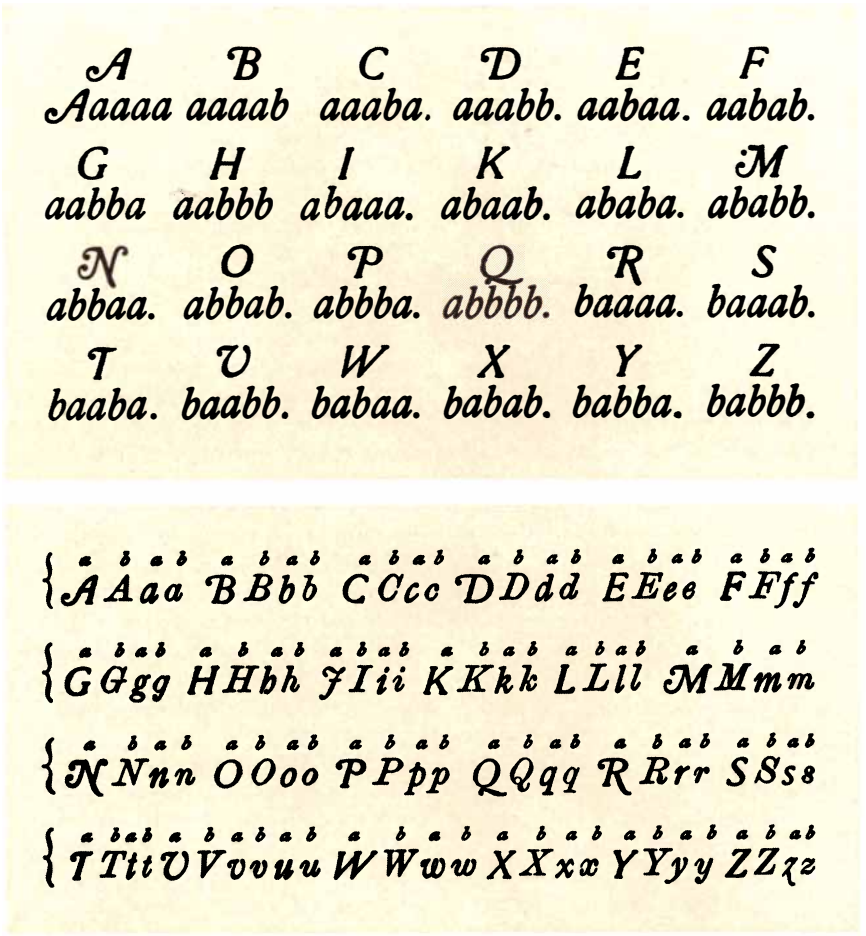
The improved loom was built in 1804. By 1812 there were 11,000 Jacquard looms at work in France. The idea also leaked out to England, notwithstanding the strained relations between the two countries, and by 1834 there were 600 Jacquard looms in Coventry. Many of them were in workers' cottages, driven by shafts and belts from common steam engines. (When one thinks of cottage industry, it is not usually in terms of a power loom in every parlor.)

It is sometimes asserted that Jacquard really invented nothing. The charge seems unfair. His change to flat surfaces on the card holder was vital. Moreover, the greatly expanded use of looms that could be reprogrammed quickly for any pattern resulted directly from his work.

Two significant developments flowed from the Jacquard loom. One was that Charles Babbage, who designed the first digital computer in the 19th century, decided that punched cards modeled on Jacquard's would be the best means of input to his machine. That was a genuine application of binary coding. Second, many of the subroutine methods and editing systems that are standard in modern computers were conceived in the 19th century to produce cards for textile patterns. Having a pattern and desiring a binary sequence for weaving it is the same as having a Fortran program and wanting the equivalent binary code that puts the program in terms suitable for a computer.

Indeed, the connection between weaving a textile fabric and designing a computing system is a close one. Anyone who looks at the wiring of a computer or at an enlargement of a large integrated circuit will notice a strong resemblance to normal fabric patterns.

The problem that Jacquard and his forerunners solved was to make widely different patterns on the same production machine, which could be changed rapidly from one pattern to another. The same problem is faced by electronics firms, which need to make each product with automatic machinery in order to achieve rapid output of identical items and also need to be able to change rapidly to a different product. Hence it is not surprising that the machines that connect computer wiring or make the artwork for integrated circuits are modeled on the Jacquard loom in the sense that they employ binary digital tape or punched cards to control the manufacturing movements.



BINARY CODE devised by Francis Bacon in the 17th century employed the letters *a* and *b*, equivalent to the 0 and 1 of modern codes, to represent the 24-letter alphabet of the time in five-letter groups (*top*). A secret message such as "Flee" would be encoded in this way. The second part of the code was a "biformed alphabet" (*bottom*) that employed two slightly different fonts of type, one representing the *a* and the other the *b*, for printing an apparently innocent message. The recipient would change the message into the *a* and *b* pattern, divide the letters into groups of five and thus reproduce the original secret message.

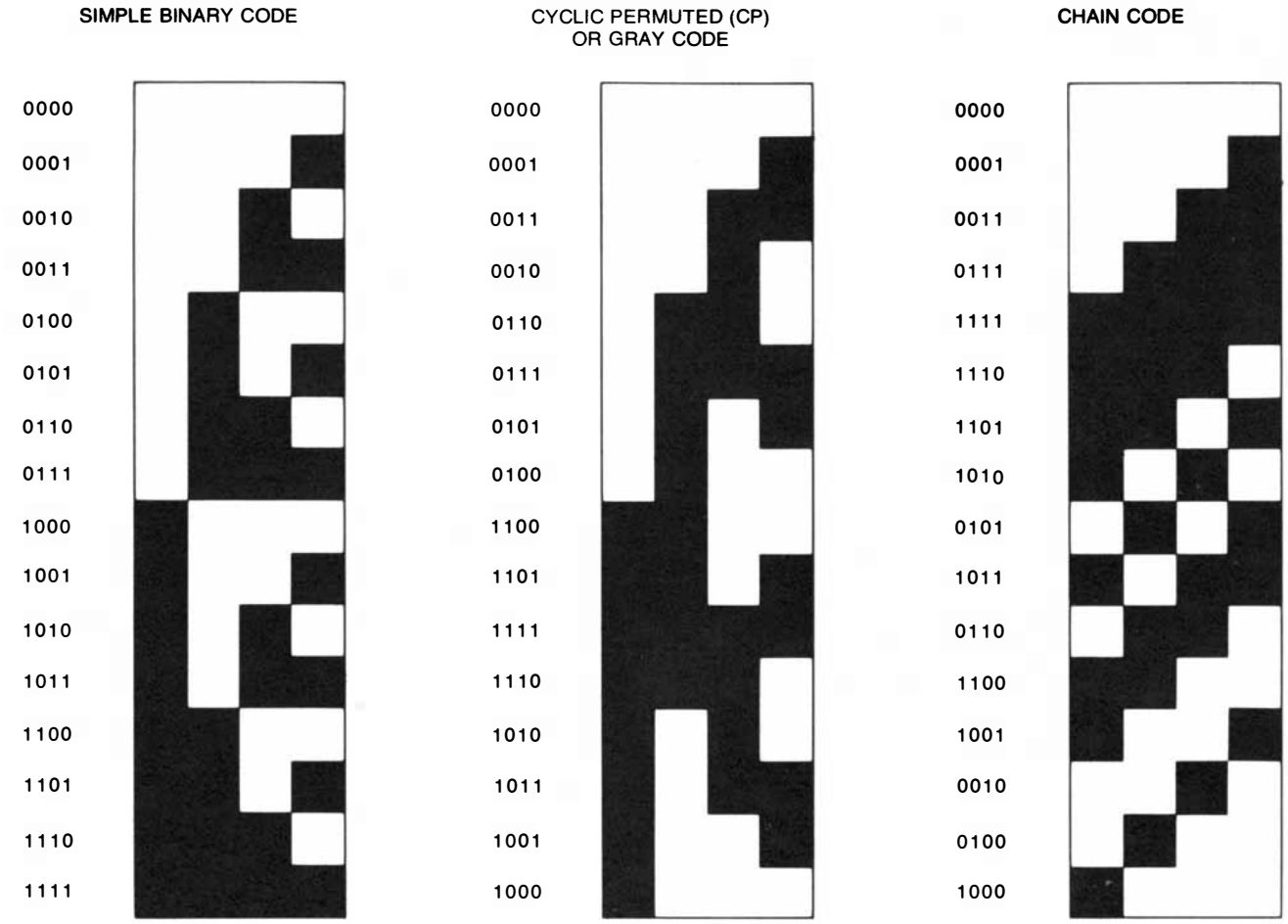
Boole's logical algebra also had forerunners, beginning with the work of the ancient Greeks, who spent a great deal of time on logic (as on geometry) and established the style and method that provided the background of European thought on the subject for centuries. Since in the Middle Ages learning was almost exclusively associated with the church, logic was used mainly in attempts to prove various theological doctrines, such as the existence of God, from first principles.

Among the classical mathematicians Gottfried Wilhelm von Leibniz came closest to discovering Boolean algebra; indeed, he identified all the processes wherein Boolean algebra does not parallel ordinary algebra. As an example, what appears to be multiplication in Boolean algebra is in fact not a multiplying process but merely a reconsideration of the same thing. For instance, if one thought of apples under the symbol A, the statement $A \times A = A$ would not represent an enlarged quantity of apples but further consideration of the subject of apples.

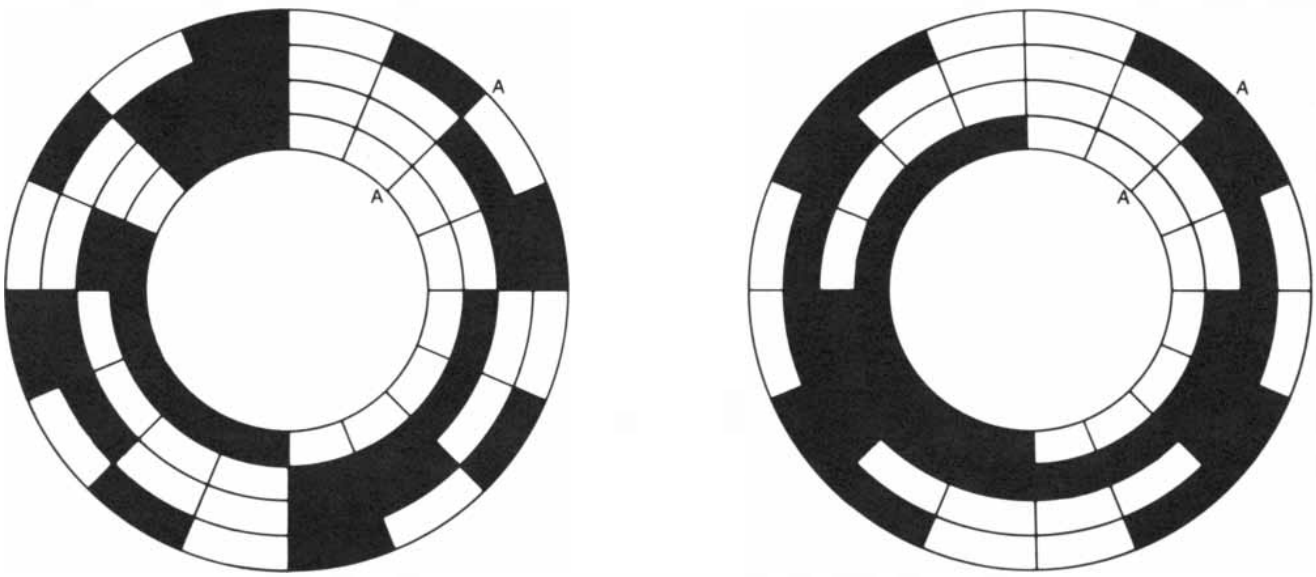
Leibniz' venture into logical algebra did not seem to excite any of his contemporaries. By the early 19th century, however, several mathematicians had taken up the subject. It remained for Boole to do the definitive work.

Boole applied algebraic symbols to typical logical arguments such as "Absolute evil is either moral evil, or it is, if not moral evil, a consequence of moral evil." He also analyzed more positive theological proofs, such as "Unchangeable and independent Being must be self-existent." The symbolic method was able to pull out many more deductions than were possible in the verbal reasoning. It also found the main weakness, namely that a particular statement might have one underlying assumption in one place and another one elsewhere.

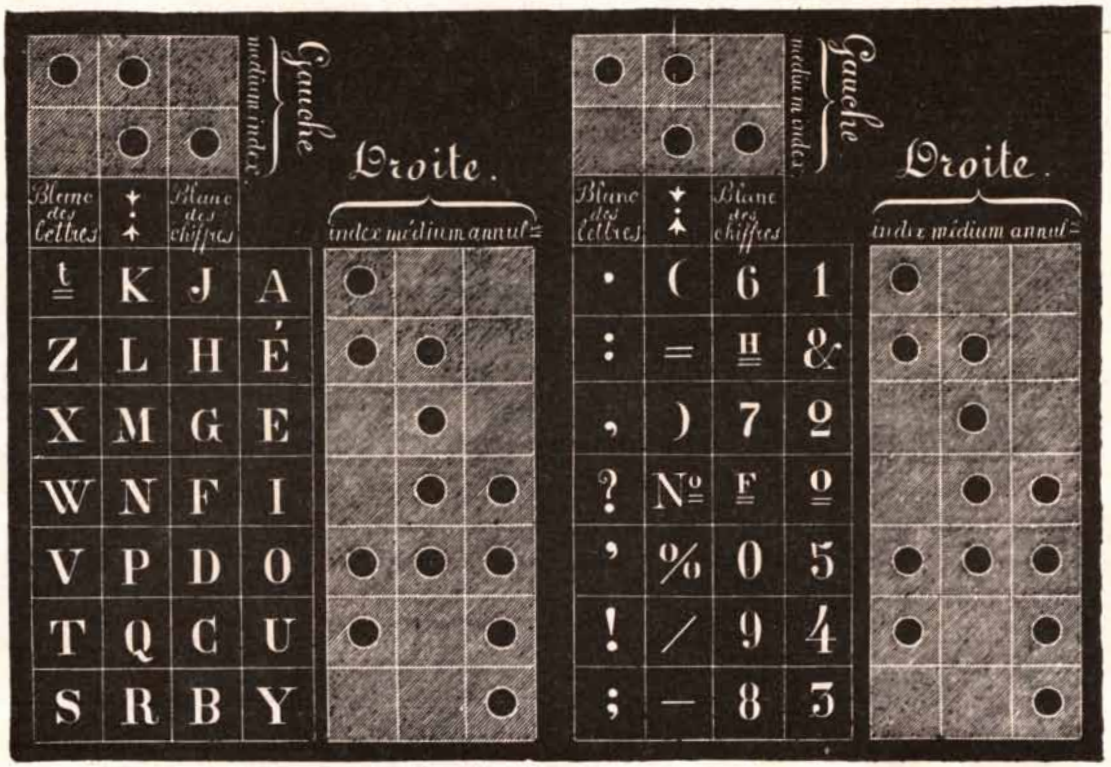
It is perhaps unfortunate that Boole's algebra—based on the functions "and," "or" and "not" applied to binary variables—is so simple, since that fact obscures Boole's great contributions to many fields of mathematics. Boole developed his algebra not in the elementary way found in computer texts but from a deep and lengthy study of thought processes and natural language. "There exist, indeed," he wrote, "certain general principles founded in the very nature of

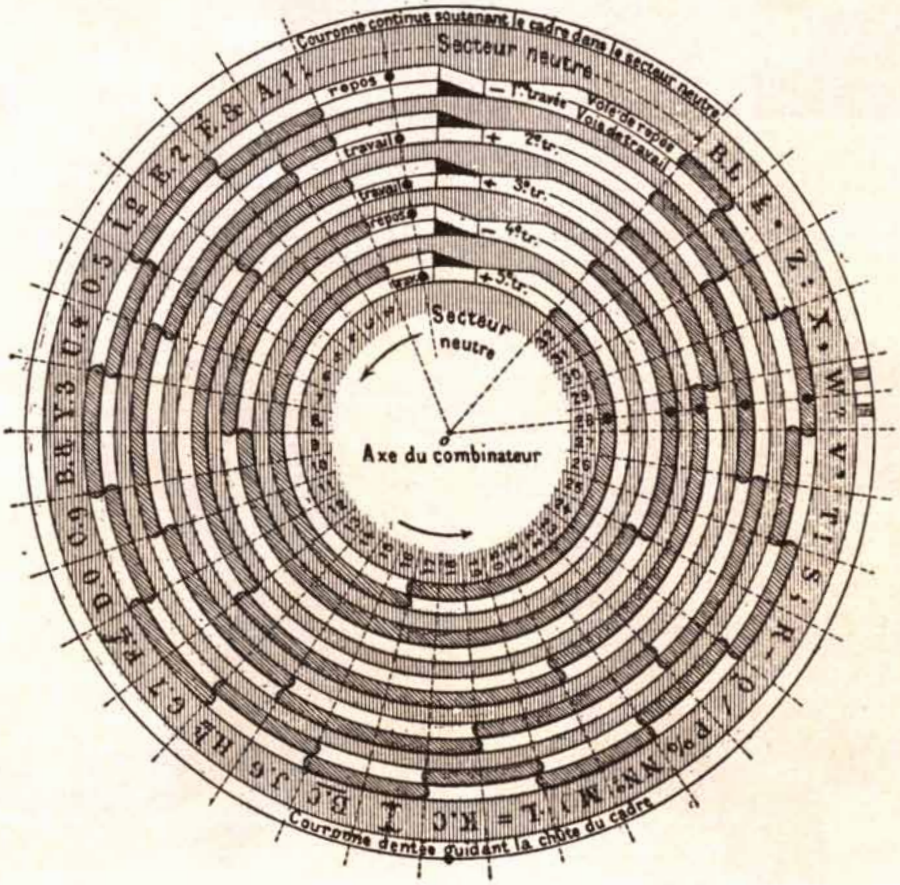


THREE BINARY CODES are depicted in a scheme employing blank spaces to represent 0 and black spaces to represent 1. A simple binary code is shown for the decimal numbers 0 through 16; the same arrangement could serve, as in telegraphy, for the letters *a* through *q*. The cyclic-permuted, or Gray, code and the chain code employ different arrangements. The Gray code reduces ambiguity in the readout, and the chain code makes decoding simpler since a single track or column reproduces the code after a one-step delay. The French engineer Emile Baudot introduced the cyclic-permuted code for telegraphy approximately 100 years ago.



READOUT DEVICES are portrayed as disks on which an insulating material (*white*) represents 0 and a metal (*black*) represents 1. Conducting brushes read the incoming data. With a simple binary code (*left*) errors in reading will occur at some of the boundaries, such as at the line *A–A*, where a brush might read 0000, 0001, 0010 or 0011. A cyclic-permuted code (*right*) changes in only one significant place at each boundary. At a boundary such as *A–A* the brush can read only a pattern adjacent to the line.





BAUDOT'S TELEGRAPH had this sending keyboard and receiving wheel. On the keyboard a blank square was 0 and a square containing a circle was 1. To transmit *F*, for example, the code was 01 for the left hand (*gauche*) and 110 for the right (*droite*). At the receiver the five binary digits were stored and the code wheel was rotated, reaching in time an angle where the pattern under the reading head was the same as the transmitted pattern. The wheel was connected directly to a printing disk, and when the received pattern appeared, the disk was actuated and the letter or number or symbol that had been transmitted by the sender was printed.

language, by which the use of symbols, which are but the elements of scientific language, is determined." The statement strikes a notably modern note.

The evolution of logical algebra since Boole's time has produced a puzzling contrast. Boole developed the few connectives of his algebra from a study of the precise use of language. Today, however, if a precise language is to be implemented by a computer, the computer must contain several thousand of these connectives, together with a compiler program occupying perhaps 8,000 words of 16 bits each. Perhaps the common interpretation of the Boolean way is not the best or the only way to process language statements.

I come now to the contributions of Baudot to printing telegraphy. It is a field in which he stands out from every point of view: for his discovery of new codes, for his means of feedback control and for the total design of his system. Before describing his work, however, I must introduce the reader to the cyclic-permuted (or Gray) code and the chain code (usually known in the U.S. as feedback shift-register sequences).

The problem faced by Baudot was to arrange for the printing in one place of letters and numerals transmitted electrically from a distant place. A similar problem arises in machine-tool control with input devices that convert a physical position, such as the angle of a shaft, into digital signals. A solution would be to make a disk on which a metal represented 1 and an insulating material represented 0 [*see bottom illustration on page 81*]. The trouble is that somewhere in the system the data must be read off by conducting brushes, and no matter how accurately the disk is made and aligned with the brushes, the output will be wrongly read across some of the boundaries. For example, a brush at the line between the position of the disk representing 0001 and the position representing 0010 could be read off as 0000, 0001, 0010 or 0011.

An improved solution is a code that changes in only one significant place at each boundary. Such a code is the cyclic-permuted, or Gray, code, which is represented by the second disk in the bottom illustration on page 81. The readout device cannot read anything other than a pattern adjacent to a transition, so that the ambiguity of the simple binary disk is removed. This property has many applications in modern automatic machines. (Other applications of the Gray code are discussed in "Mathematical Games," beginning on page 106.) The cyclic code still bears a direct relation to the binary, as can be seen from the illustration.

The chain code is a different matter. It has no digit-by-digit relation to binary; each chain code is its own unique numbering system and can only be correlated with binary by counting along the code [*see top illustration on page 81*]. The code has several properties that are valuable in electronics and computer science. In a four-digit pattern the three digits on the right side of one pattern are the three digits on the left side of the next-lowest pattern. Hence the codes can be generated in a simple shift register. Moreover, each column, after a delay of one step, is the same as the one on its right, so that a single column or track represents the code.

If more digits are put into the code, the code acquires the property of detecting and correcting errors. An example may be illuminating. A simple four-bit, 14-pattern code is 11110010000110/ (repeats). This is a skew-symmetric code, where the second half is an inversion of the first half, that is, the first half consists of four 1's, two 0's and a 1 and the second half consists of four 0's, two 1's and a 0.

If four-bit patterns are designated as defining position in the code, one bit in error gives a wrong result. For example, 1100 would change to 0100, 1000, 1110 or 1101, all of which are valid patterns in the code. If six bits are used to define position, however, an error will be detected if one digit is wrong. For example, 110010 would change to 010010, 100010, 111010, 110110, 110000 or 110011. None of these is present in the original code sequence, so that the error would be detected.

A seven-bit pattern can correct a single error. The reason is that a single change in a seven-bit pattern produces a corrupted version that is closer in structure to the original pattern than any alternative actually present in the original set of patterns. The reader should try this with, say, 1100100. A few trials with pencil and paper will prove the assertion.

Baudot's first telegraph had a transmitter keyboard that employed five binary digits to transmit each letter, number and symbol [*see illustration on opposite page*]. The keyboard had one case for letters and one for numbers and symbols, rather than the lower case and upper case of modern typewriters. The code was cyclic-permuted.

In the receiver the five binary digits were stored on a code wheel. The wheel was connected directly to the printing disk. As the disk rotated it would eventually reach an angle where the pattern under the reading head was identical with the received pattern, whereupon the printing head was actuated and printed the letter, number or symbol that had been transmitted.

An independent reference states that Baudot devised the code for its property of minimum commutation, showing that he understood how to find specific codes from a definition of desirable properties. His talents were immediately recognized after he had exhibited his equipment at the Universal Exposition in Paris in 1878. He received a gold medal (two others went to Thomas A. Edison and Alexander Graham Bell) and was made a chevalier of the Legion of Honor.

Baudot's printing telegraph went on to become the standard item. In about 1882 he redesigned it on the coding side so that it operated on a chain code. (The date is uncertain because the equipment was not shown until the Universal Exposition of 1889 in Paris, and Baudot did not write about it until 1895.) An extract from his description of the system shows that once again he had conceived of a useful code property (only one track instead of five tracks) and had taken steps to find it. "From the point of view of construction," he wrote, "it is interesting to find, among the milliards of arrangements possible, those able to give 31 patterns an order which allows one to arrange the flanges and the spaces in the same way on each track of a code wheel. For these researches and many others, I have been effectively helped by M. Cartier, Chief Research Engineer of the Carpentier workshops, whose ingenuity and modesty are well known to all those inventors who have worked with him."

The historian of science is not too pleased about the modesty of M. Cartier. It would be interesting to be able to read about the approach he took to the problem before the days of shift registers and applied Boolean algebra. There is no doubt, however, that Baudot was the dominant figure. Indeed, he can be regarded as the inventor of both the cyclic-permuted code and the chain code.

Binary codes have of course proliferated since the first binary computers were designed and built starting in about 1949. It is relatively easy to follow the course of discovery since that time. What emerges from such an effort is the fact that in general the applications only represent refinements of properties found in the much older codes that I have described.

The Neurophysiology of Binocular Vision

The ability of certain mammals, including man, to visually locate objects in the third dimension is traced to the selective activity of single binocular nerve cells in the visual cortex of the brain

by John D. Pettigrew

Man, along with cats, predatory birds and most other primates, is endowed with binocular vision. That is to say, both of his eyes look in the same direction and their visual fields (each about 170 degrees) overlap to a considerable extent. In contrast, many animals, such as rabbits, pigeons and chameleons, have their eyes placed so as to look in different directions, thereby providing a more panoramic field of view. Two questions come to mind: First, why do we have binocular vision instead of panoramic vision? Second, how is it that our single impression of the outside world results from the two different views we have of it by virtue of the separation of our eyes?

In answer to the first question, it is now known that binocular vision provides a powerful and accurate means of locating objects in space, a visual aptitude called stereopsis, or solid vision. Of course, it is possible to judge distance from the visual image of one eye by using indirect cues such as the angle subtended by an object of known size, the effort used in focusing the lens of the eye or the effect of motion parallax (in which the relative motions of near and far objects differ). These cues cannot be used in all situations, however, and they are not as accurate or as immediate as the powerful sensation of stereopsis, which is perhaps most familiar in the context of stereoscopic slide-viewers, three-dimensional motion pictures and so on. Some 2 percent of the population cannot enjoy stereopsis because of undefined anomalies of binocular vision. It is the aim of this article to give an account of recent work that shows how the brain achieves the very first stages of binocular depth discrimination.

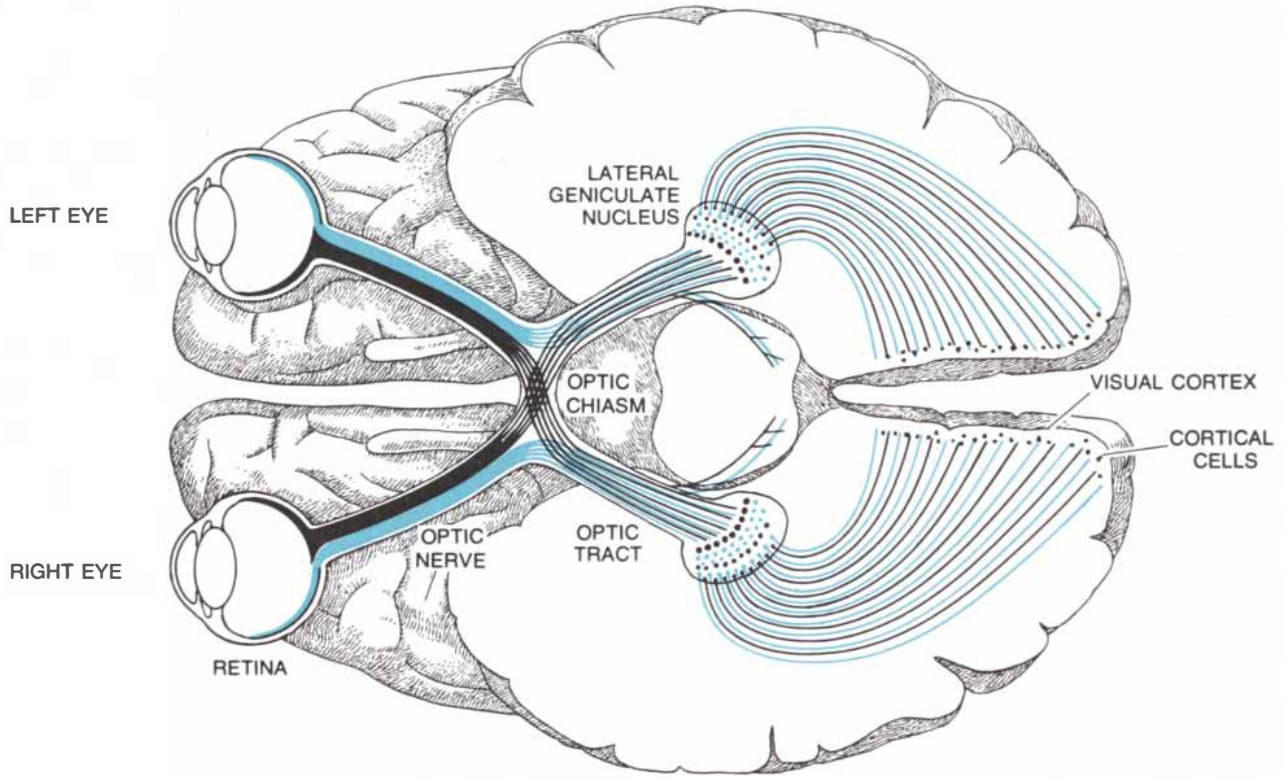
Although it was not until the 19th century that the advantages of binocular vision were clearly demonstrated, man has pondered the arrangement of his eyes from earliest times. Of more concern to early investigators was not the first question, "Why binocular vision?" but rather "How does my single unified impression of the world result from the two views I have of it?" This second question is almost as difficult to answer today as it was when it was first asked by the ancient Greeks. The problem of "fusing" two slightly differing views of the world, however, is closely akin to the problem of using the slight differences between the views to achieve stereopsis. Thus a better understanding of the events in the nervous system underlying stereopsis should also throw some light on the problem of binocular fusion.

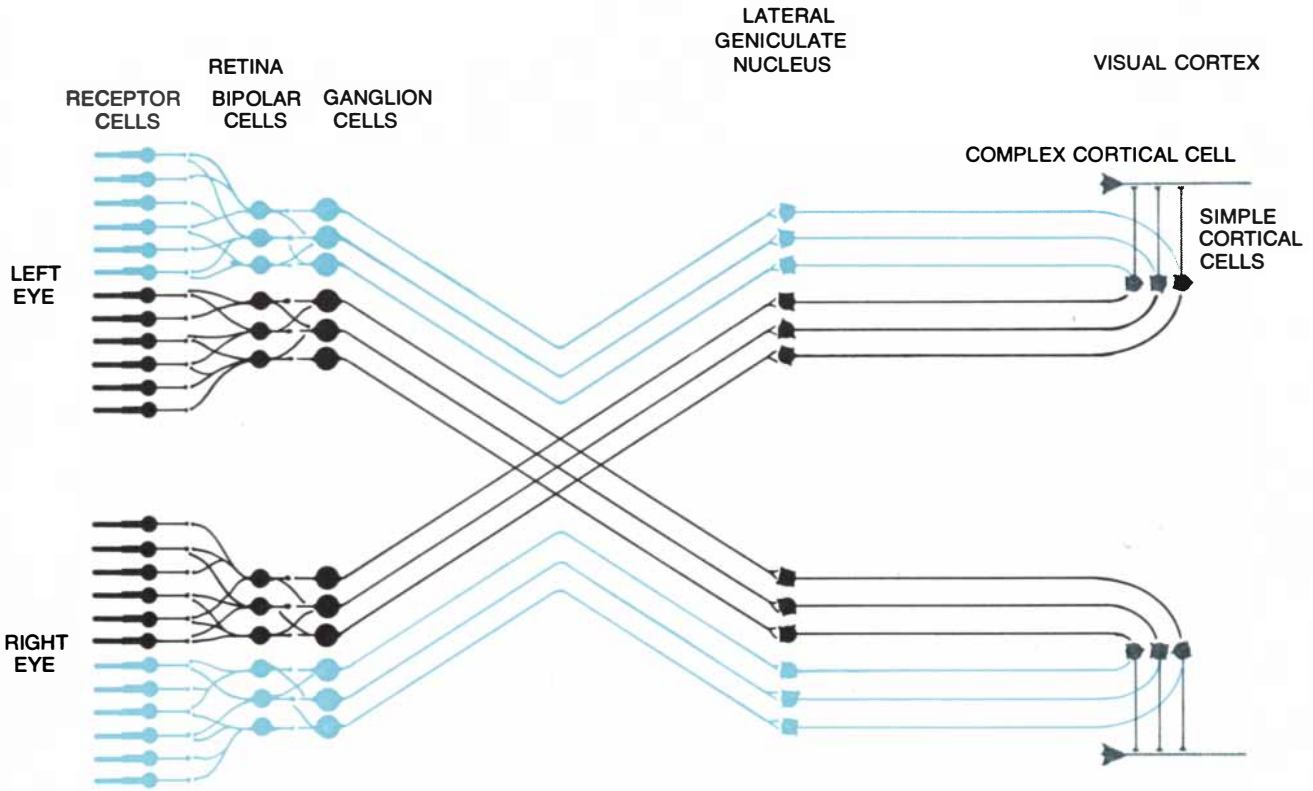
Galen taught in the second century that the fluid-filled ventricles of the brain were the seat of union, with a flow of visual spirit outward to both eyes. Galen's teachings were influential until the Renaissance, when scholars realized that the transfer of information is from the world to the eye, rather than in the reverse direction. René Descartes proposed in the 17th century that fibers from each eye might converge on the pineal gland for unification [*see illustration on page 86*]. His scheme, although incorrect, clearly indicates the now established principle that fibers from roughly corresponding regions of each eye converge on a single site in the brain. It was Isaac Newton who in 1704 first proposed that where the optic nerves cross in the optic chiasm there is an exchange of fibers. An early drawing of this concept, called partial decussation, shows how the fibers come together to carry information from corresponding parts of each eye [*see illustration on page 87*].

Newton's proposal, unlike that of Galen or Descartes, has been extensively verified. The number of uncrossed fibers in the optic chiasm depends on the amount of overlap of the two visual fields, and this number tended to increase as animals evolved with eyes occupying a more frontal position [*see illustration on page 88*]. The rabbit, with only a tiny binocular portion in its visual field, has a very small number of ipsilateral, or uncrossed, fibers in the optic chiasm, whereas each of its cerebral hemispheres is heavily dominated by contralateral, or crossed, fibers from the opposite side. As the amount of binocular overlap increases from animal to animal, so does the number of ipsilateral fibers. In man there is almost complete overlap and 50 percent of the fibers of the optic nerve are uncrossed.

Although partial decussation provides the opportunity for the optic nerve fibers to come together in the brain, for a long time there was controversy over whether this coming together does in fact occur. For instance, at the first way station for the optic-nerve fibers in the brain, the lateral geniculate nucleus, the inputs from the two eyes are carefully segregated into layers. The more binocular overlap the animal has, the more obvious the layering is. The segregation is confirmed by physiological recordings that show that a neuron, or nerve cell, in a given layer can be excited by light stimuli falling on one eye only. The segregation is reinforced by inhibitory connections between corresponding neurons in adjacent layers.

At the level of the visual cortex of the brain, however, single neurons do receive excitatory inputs from both eyes. David H. Hubel and Torsten N. Wiesel of the Harvard Medical School demonstrated this effect for the first time in





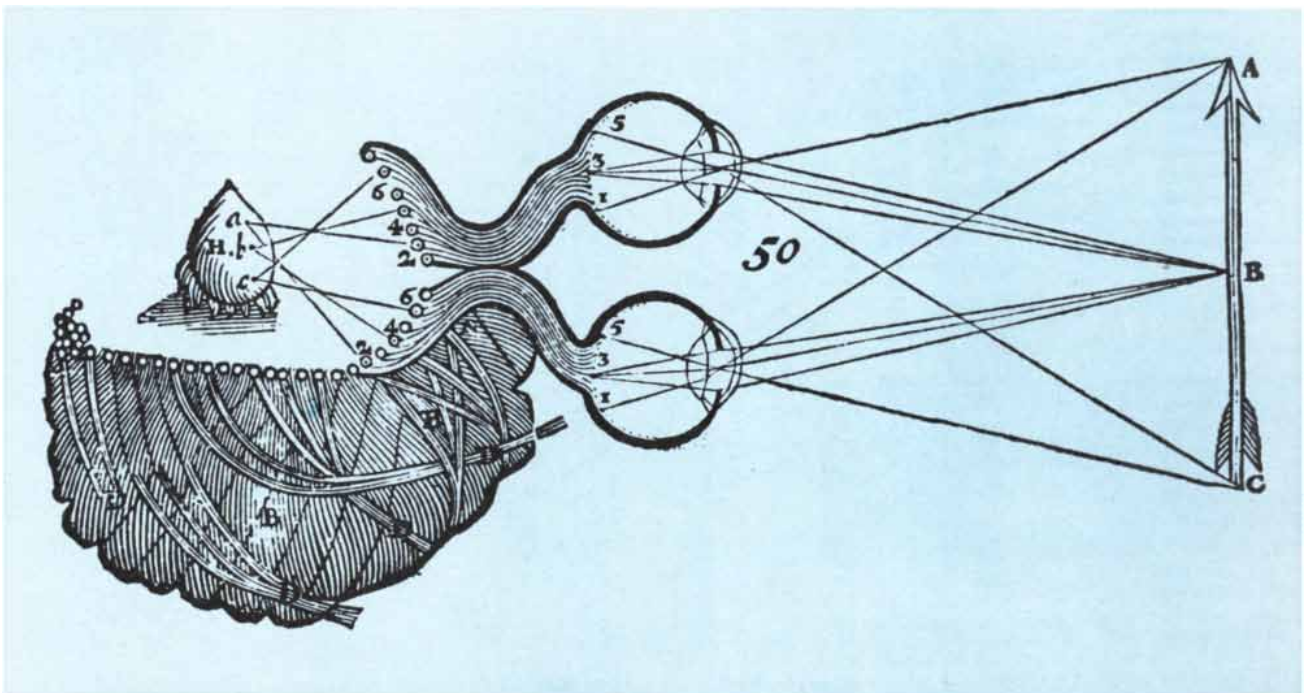
ANATOMY OF BINOCULAR VISION is represented in the drawing at top, which shows the human brain as viewed from below. The visual pathway from retina to cortex consists essentially of six types of neurons, or nerve cells, of which three are in the retina, one is in the lateral geniculate nucleus and two are in the cortex (*see schematic diagram at bottom*). Roughly half of the fibers of the optic nerve from each eye remain uncrossed at the optic chiasm. These ipsilateral, or uncrossed, fibers (*color*) from the outer part of the retina of one eye join with the contralateral, or crossed, fibers (*black*) from the inner part of the retina of the other eye, and the two types of fiber travel together along the optic tract to the lateral geniculate nucleus, where fibers from each eye are segregated into layers. The fibers that emerge from this body carrying an input from both eyes converge on single neurons in the cortex.

1959 by recording from single neurons in the visual cortex of the cat [see "The Visual Cortex of the Brain," by David H. Hubel; SCIENTIFIC AMERICAN, November, 1963]. For almost every nerve cell studied two areas could be defined where light stimuli evoked a response, one associated with each eye. The proportion of binocularly activated neurons found has increased as the technique of presenting stimuli has become more refined. Recent work by P. O. Bishop, Geoffrey H. Henry and John S. Coombs of the Australian National University shows that all the cells in the striate cortex of the cat receive an excitatory input from both eyes. Since each neuron in the striate cortex is simultaneously "looking" in two directions (one direction through each eye), the striate cortex can be regarded as the "cyclopean eye" of the binocular animal. The term cyclopean eye, derived from the mythological Cyclops, who had one eye in the center of his forehead, was first used by Ewald Hering and Hermann von Helmholtz in the 19th century to describe the way the visual cortex resolves the different directions of a given object as seen by each eye. Besides assessing visual direction, the cyclopean eye has an ability not possessed by either eye alone: it can use binocular parallax to ascertain the distance of an object.

Imagine looking down on an upturned bucket [*see illustration on page 89*]. If one directs each eye's fovea (the central high-resolution area of the retina) toward the cross at the center of the bucket, each eye will receive a slightly different view. The difference between the two views is called binocular parallax. Since each retinal image of the bottom of the bucket (small circles) is equidistant from the fovea, these two images must lie on exactly corresponding retinal points and are said to have zero disparity. Because of the horizontal separation between the eyes, the images of the rim of the bucket (large circles) are displaced horizontally with respect to the small circles and the foveas. These images lie on disparate retinal points. The retinal disparity between two such images is measured as an angle that corresponds to the difference between the angular separations of the two images from some known point such as the fovea (in the case of absolute disparity) or the smaller circle (in the case of relative disparity). If the disparity between the retinal images of the large circle is not too great, one sees not two large circles but a single large circle floating in depth behind the small one.

The basis of this powerful depth sensation of stereopsis was first demonstrated by Sir Charles Wheatstone (of Wheatstone-bridge fame) in 1838. By providing very precise localization of objects in visual space, stereopsis can be regarded as the *raison d'être* of binocular vision. Whenever in evolution the need for the protection of panoramic vision was lessened (by the animal's taking to the trees as in the case of the primates, or by the animal's becoming predatory as in the case of the cats), then binocular vision developed to make it possible to use a depth cue more direct and accurate than the depth cues available to one eye alone. In addition, stereopsis enables a predator to penetrate the camouflage used by its prey, because monocular form perception is not a necessary prerequisite for stereoscopic vision. For example, an insect disguised as a leaf may be invisible monocularly but stand out in a different depth plane from real leaves when it is viewed stereoscopically. One can readily demonstrate this effect for oneself with the aid of random-dot stereograms devised by Bela Julesz of the Bell Telephone Laboratories [see "Texture and Visual Perception," by Bela Julesz; SCIENTIFIC AMERICAN, February, 1965]. Here a given pattern, such as a square, may be invisible to monocular inspection but stand out vividly when viewed stereoscopically.

The sole basis of stereopsis is the horizontal disparity between the two retinal

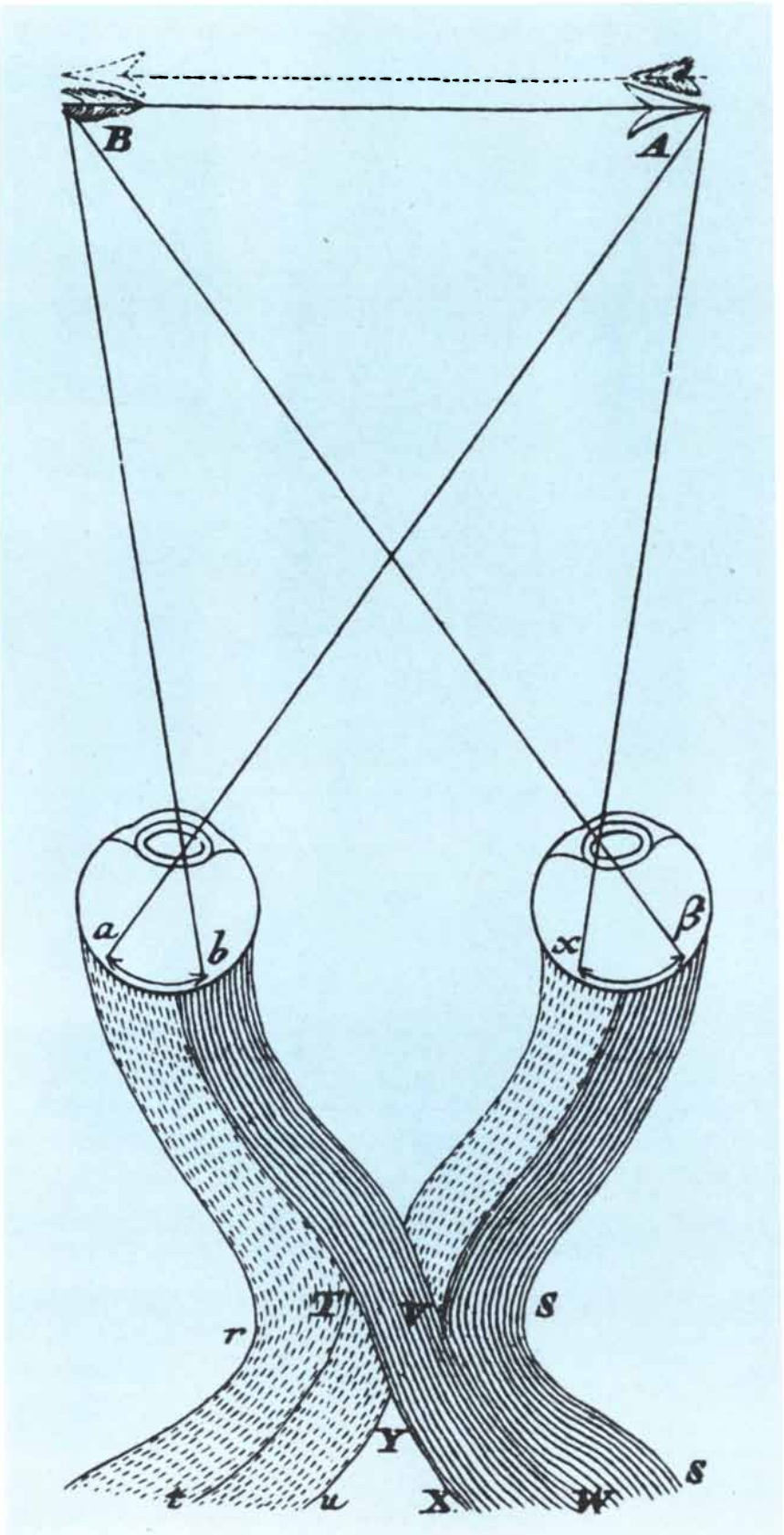


CONVERGENCE of nerve fibers from corresponding regions of each eye on a single site in the brain was proposed more than 300 years ago by René Descartes. In this early drawing of Descartes's scheme, reproduced from his study *Traite de l'Homme*, the optic-nerve fibers from each eye are shown converging on the pineal gland (*H*), where they are rearranged, with those from corresponding retinal regions merging together. It is now known that fibers from both eyes do in fact converge, but not on the pineal.

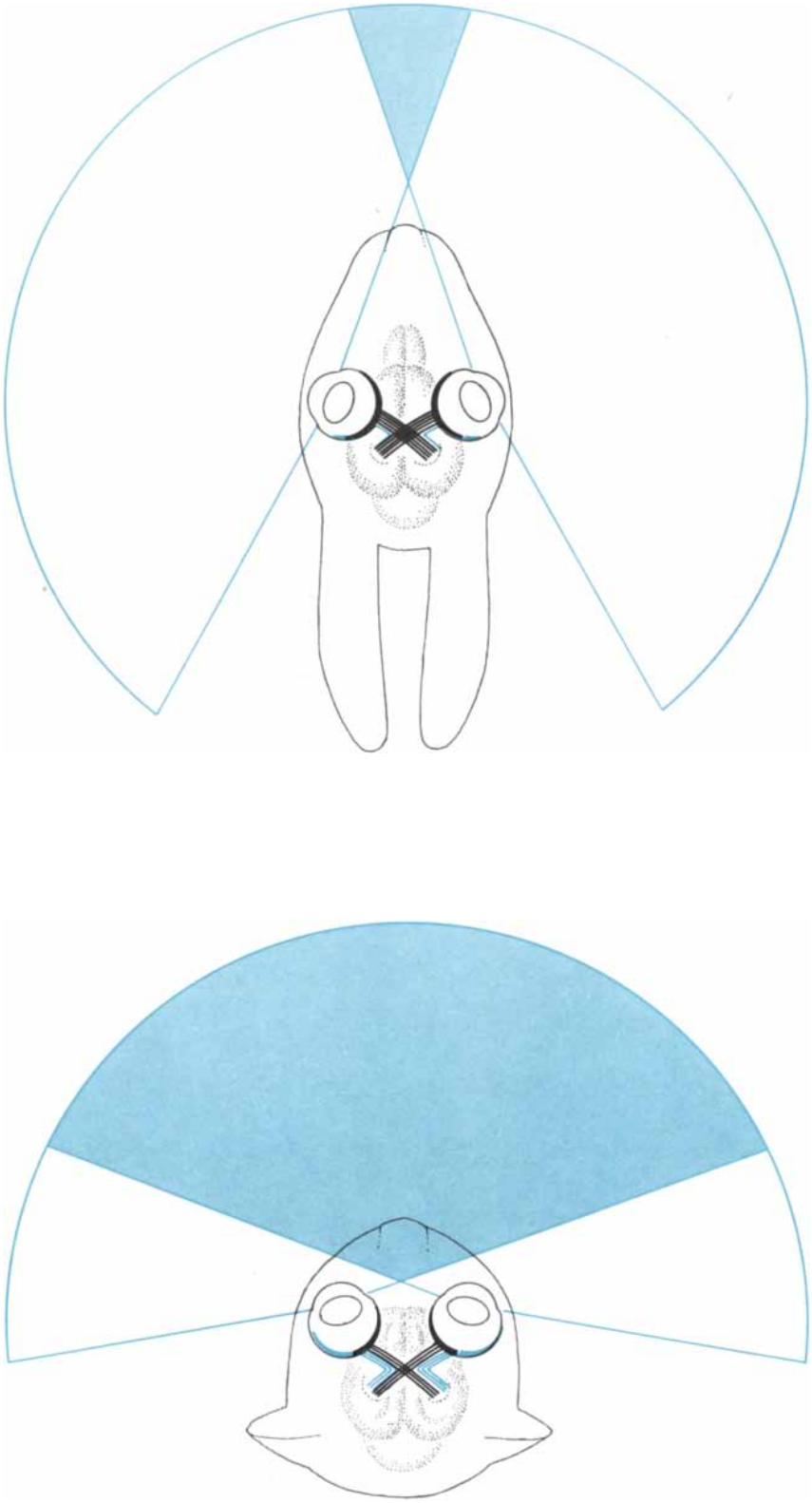
images. Of course, the brain must have a means by which it can first select those parts of the two images that belong to each other in the sense that they are images of the same feature in space. The horizontal disparities between the paired parts give the cue to depth. Julesz' experiments show that the binocular assessment of depth does not require the prior recognition of form, suggesting that the disparity information is processed by the brain fairly early in visual perception. Encouraged by this suggestion, neurophysiologists have examined the properties of binocularly activated neurons in the visual cortex, and over the past five years they have learned more about the neural mechanisms of binocular depth discrimination.

The experimental arrangement for studying binocularly activated neurons in the visual cortex is shown in the illustration on page 90. A cat that has been anesthetized and given a muscular paralyzing agent to prevent eye movements faces a screen onto which a variety of visual stimuli can be projected. A microelectrode inserted into the visual cortex samples activity from single nerve cells, and this activity is amplified so that it can be displayed on an oscilloscope, recorded and fed into a loudspeaker, in order that the experimenter can readily follow the response. Unlike neurons in the retina or the lateral geniculate nucleus, cortical neurons need exquisitely defined stimuli if they are to fire. Each cell requires that a line of a particular orientation (for instance a white slit on a dark background, a dark bar on a light background or a dark-light border) be placed on a narrowly defined region of each retina. Neurons with the same requirements for orientation of the stimulus are grouped together in columns that run from the surface of the brain to the brain's white matter. For some neurons within the column absolute position of the line is very important. Plotting with small spots of light suggests that such neurons receive a fairly direct input from the fibers coming into the cortex. These are "simple" neurons. Other neurons, called "complex," probably signal the output of the column; they behave as if they receive an input from a large number of simple neurons.

The speed and the direction of the moving stimulus are also important. Each of these stimulus requirements is more or less critical and each varies from neuron to neuron. As one slowly advances the electrode to pick up neurons,



PARTIAL DECUSSATION, or crossover, of nerve fibers at the optic chiasm was first proposed by Isaac Newton in 1704. This drawing, made by a contemporary of Newton's, shows how fibers carrying information from corresponding parts of the eye come together on the same side of the brain in the interests of binocular vision. Newton's scheme was verified in the 19th century by the first ophthalmologists, Heinrich Müller and Bernhard von Gudden.



AMOUNT OF BINOCULAR OVERLAP of an animal's two visual fields is proportional to the percentage of uncrossed fibers in the optic chiasm. As animals evolved with eyes occupying a more frontal position this percentage tended to increase. The rabbit, for example, has only a tiny amount of binocular overlap and accordingly has a very small number of uncrossed fibers (*top*). The cat, in contrast, has a much larger binocular overlap and a correspondingly higher percentage of uncrossed fibers (*bottom*). In man there is almost complete binocular overlap and 50 percent of the fibers are uncrossed. The uncrossed nerve fibers carry information from the outer part of retina, the region responsible for binocular vision.

one must be continually moving a complicated pattern in front of the animal's eyes, in order to activate neurons that would otherwise be missed because of their lack of activity in the absence of the specific stimulus. (I sometimes wear a knitted sweater with a regular design as I move about in front of the cat.) Once one has found the specific stimulus for a given neuron, it is possible to define a region in the visual field of each eye where that stimulus will cause excitation of the cell. This region, called the response field, is plotted for each eye on a screen in front of the animal. The eye on the same side of the brain as the neuron in question is called ipsilateral; the eye on the opposite side is called contralateral. The study of a number of neurons in succession gives an array of ipsilateral and contralateral response fields [*see illustration on page 92*]. The arrays for each eye are separated on the screen because of the slightly divergent position the eyes assume in paralysis. Normally the eyes would be lined up so that the arrays could overlap and a single object might stimulate both response fields for a given neuron.

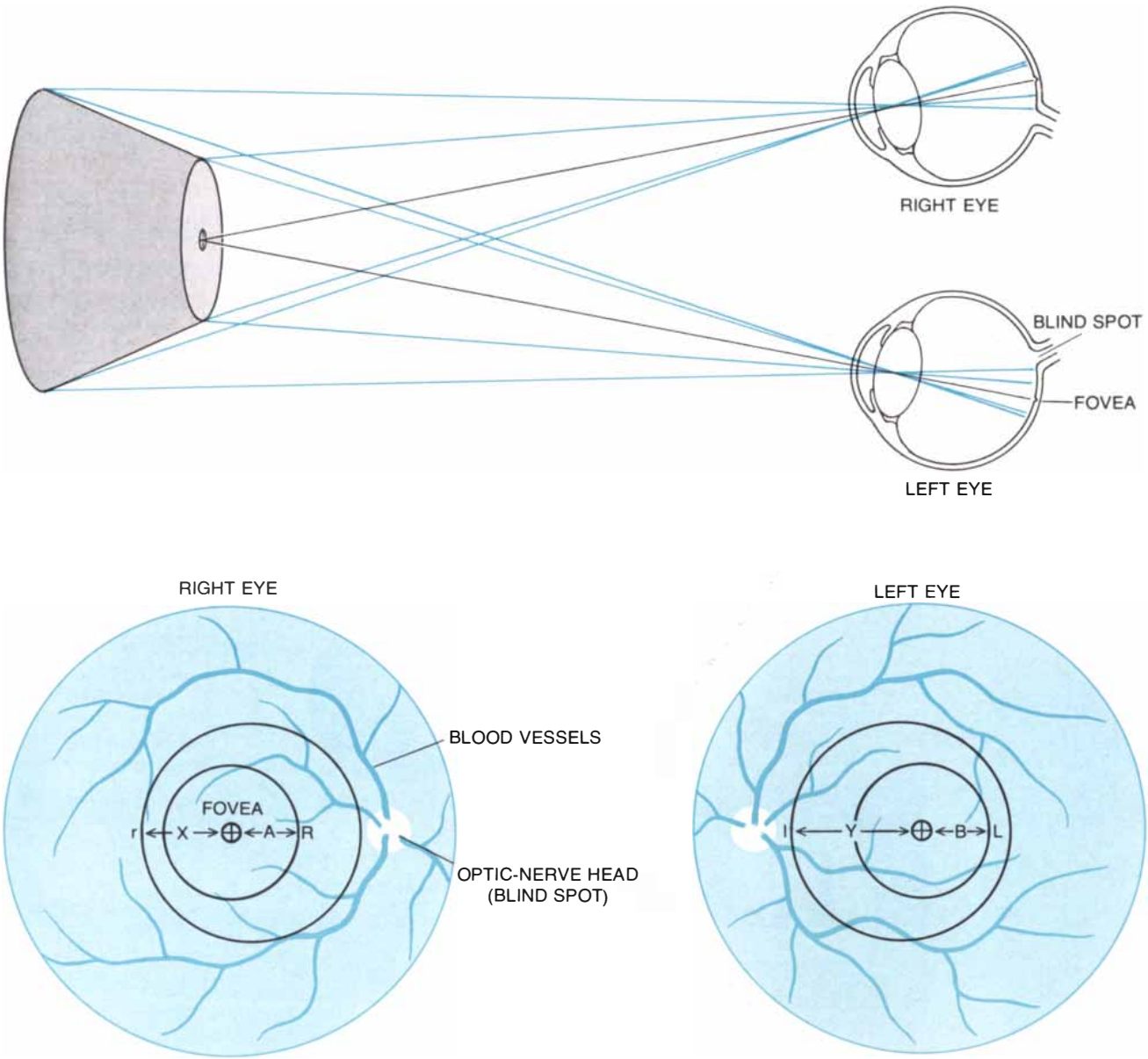
The highly specific stimulus requirements of cortical neurons could provide a means of identifying the parts of the two images corresponding to a single feature in object space. Because the number of identical features in a small part of the image in one eye is likely to be low, similar features lying in roughly corresponding regions in each eye can be assumed to belong to the same object. For example, a black line with a particular orientation and direction of movement in one image would be associated with a similar line at the most nearly corresponding position in the other image, because both are likely to be images of the same object. Since binocular cortical neurons have properties suited to the detection of the pair of retinal images produced by a given object, it was of great interest to see if they could also detect disparity between the pairs of images.

When one takes a close look at the position of each response field compared with the position of its partner in the opposite eye, it is immediately obvious that it is not possible to superimpose every response field on its partner simultaneously because of the greater scatter in the fields of the ipsilateral eye compared with those of the contralateral eye. The response fields therefore do not lie in corresponding regions of each retina and may be said to show disparity. I had

noticed this phenomenon in 1965 while working with Bishop at the University of Sydney and had considered the possibility that the variation in the position of one eye's response field with respect to the position of the corresponding response field of the opposite eye might play a role in the detection of retinal-image disparity and therefore in binocular depth discrimination. At that time, however, there were two major difficulties involved in the interpretation of the phenomenon.

The first difficulty was residual eye movement, which is present in small amounts even after the standard muscular paralyzing agents are applied. Since determination of the two response fields for one neuron can take hours (because one has to find the best stimulus orientation, speed of movement, exact position on the screen and so forth), one has to be sure that the eyes do not move in that time. Eye movement would produce spurious response-field disparities.

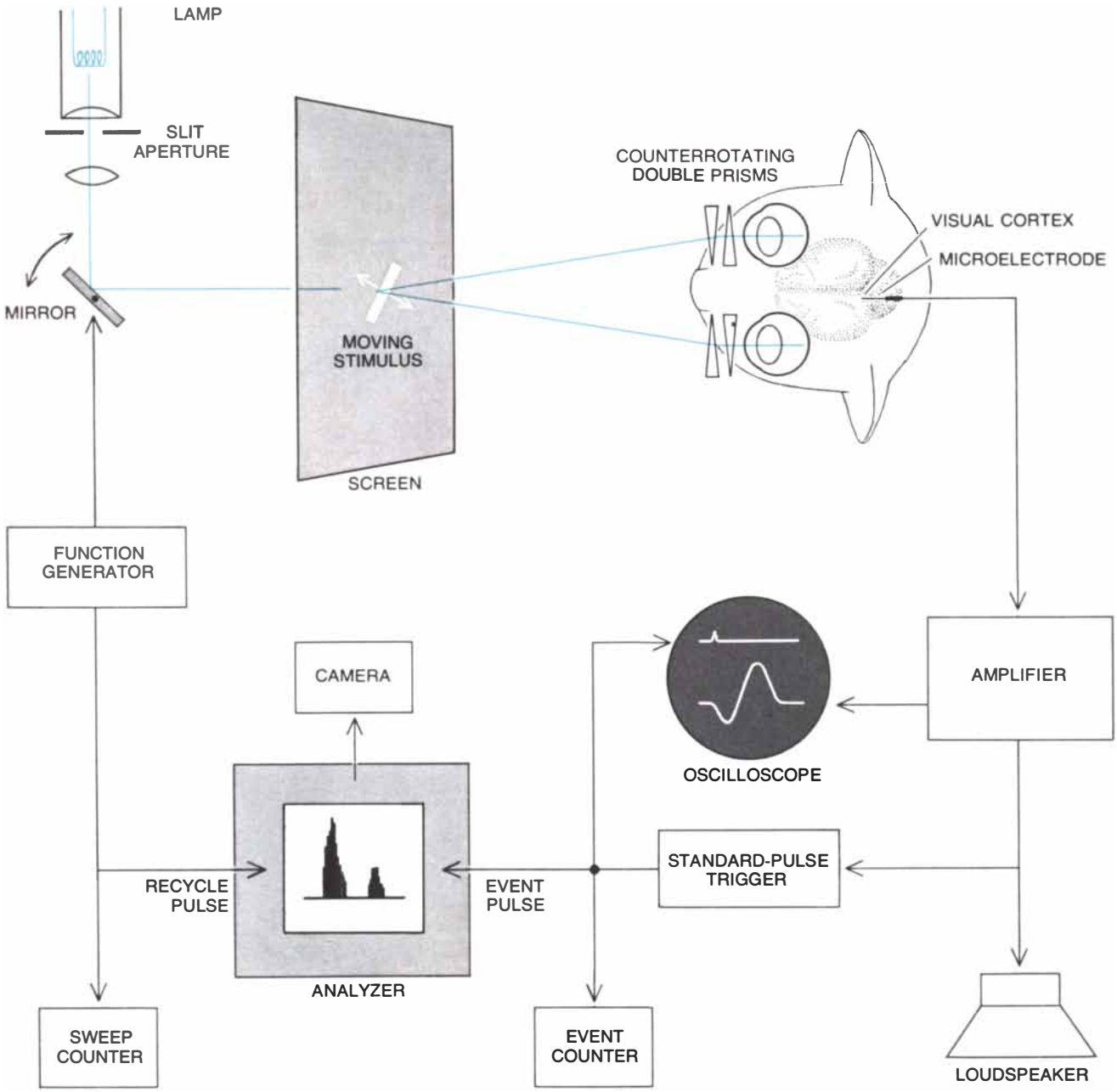
The second problem concerns the specificity of a neuron to binocular stimulation in the situation where a single stimulus is presented simultaneously to both eyes. It could be argued that the response-field disparities observed are not significant functionally since the neuron might tolerate large amounts of



BINOCULAR PARALLAX is the term used to describe the disparity produced between two retinal images when one views a three-dimensional object. In this case a subject has been instructed to direct each eye toward the cross at the center of an upturned bucket (*top*). The drawings at bottom show how the fundus of each eye would look in an indirect ophthalmoscope. The fovea, the central high-resolution area of each retina, would appear as a shallow depression on which the cross would be imaged. The optic-nerve head, or blind spot, where the fibers of the optic nerve leave the retina and the blood vessels enter, would appear as a white disk. Since the small circles corresponding to the retinal images of the bottom of the bucket are each equidistant from the fovea (that is, distance A is equal to distance B), these two images must lie on exactly corresponding retinal points (R and L); these images are said to have zero disparity. The large circles corresponding to the images of the rim of the bucket, on the other hand, are displaced horizontally with respect to the small circles and the foveas (that is, distance X is not equal to distance Y); these images are said to lie on disparate retinal points (r and l). The amount of retinal disparity, usually expressed as an angle, is the difference between X and Y.

overlap of its receptive fields without changing its response. If the tolerated overlap were of the same order of magnitude as the variation in response-field disparity from neuron to neuron, then the latter variation would be of no use.

Both of these problems were worked out in the succeeding years by me in collaboration with Bishop and Tosaku Nikara at Sydney, and with Horace B. Barlow and Colin Blakemore at the University of California at Berkeley. The problem of eye movement was solved by resorting to a number of measures simultaneously. A particularly potent mixture of neuromuscular blocking drugs was developed to reduce eye movement to a minimum without toxicity to the cat's heart and blood vessels. The sympathetic nerves to the orbit of the eye were cut to eliminate movements due to the involuntary muscles near the eye. Any residual drift was carefully monitored by plotting the projection of some small blood vessel inside the eye onto a screen. Any tiny amount of movement between response-field plots could then be corrected. In the Berkeley experiments we carefully attached the margins of each eye to rigidly held rings, which kept the eyes fixed for the duration of the long measurements.



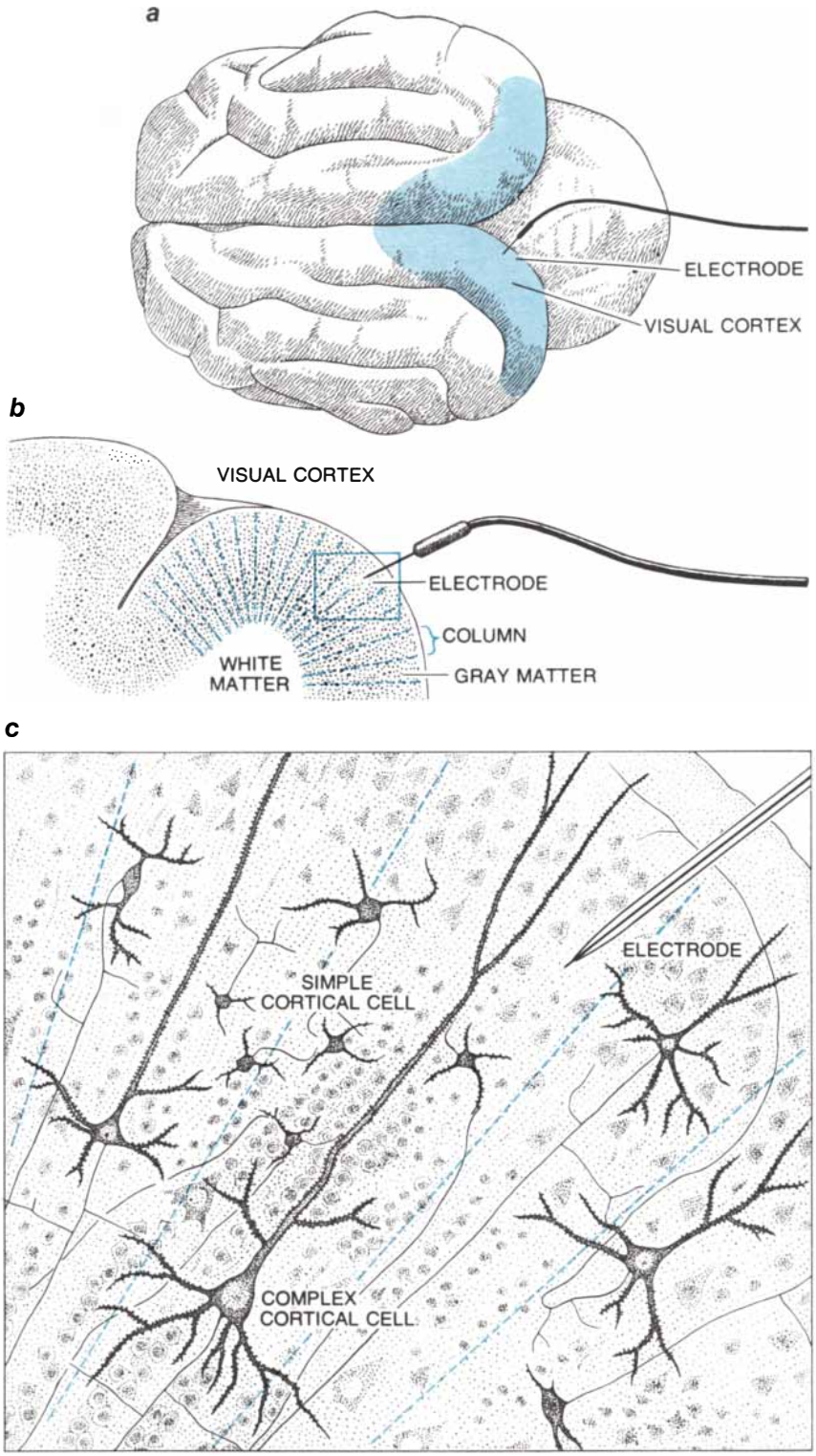
EXPERIMENTAL SETUP used by the author and his colleagues to study binocularly activated neurons in the visual cortex of the cat is depicted here. A cat that has been immobilized by a number of measures faces a screen onto which a moving line stimulus of any orientation, direction and speed can be projected from behind. A microelectrode inserted into the visual cortex samples activity from single neurons, and this activity is amplified so that it can be displayed on an oscilloscope, recorded and fed into a loudspeaker. Once a particular orientation and direction are discovered that will make the neuron fire, the stimulus is moved back and forth repeatedly while the neuron's response pattern is worked out. Counterrotating double prisms of variable power placed before the eyes enable the experimenter to determine the effect of changing retinal disparity as the stimulus moves in a fixed plane in front of the cat.

The question "How specific is the response of a single neuron for the retinal disparity produced by a stimulus presented to both eyes?" was answered with the help of the following technique. The specific stimulus was swept forward and backward over both of the response fields of a cortical neuron when these had been lined up on the screen by the use of a double prism of variable power. The double prism was used to vary the visual direction of one eye (and therefore any response fields of that eye) in finely graded steps. Since the eyes are fixed, this maneuver changes the disparity between the retinal images of the moving stimulus, and if the change in prism power for one eye is in the horizontal plane, then the effect is identical with the one produced by a change in the distance of the stimulus along a line through the opposite eye and its response field.

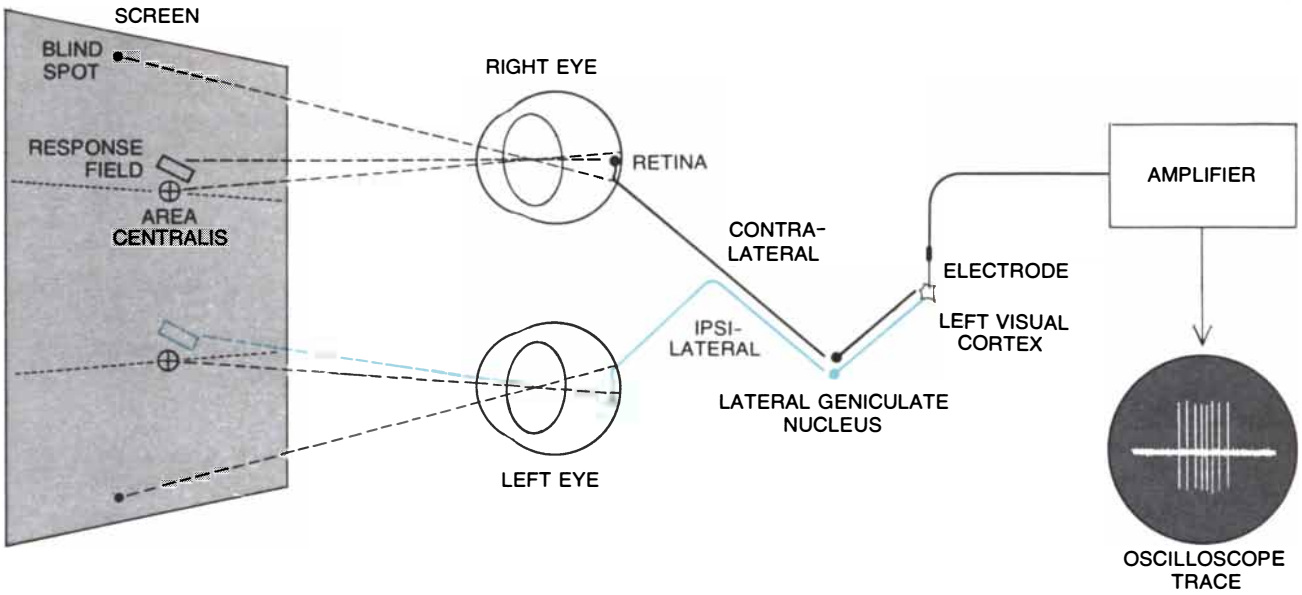
We found that the binocular response of a particular cortical neuron in the left visual cortex to a moving slit of light varies considerably as the disparity is changed [*see illustration on page 94*]. The responses were averaged for stimulation of each eye alone and for binocular stimulation when the two response fields have different amounts of overlap on the tangent screen where the light stimulus is moving forward and backward. The cortical neuron responded only to an oblique slit, and there was no response when the orientation was rotated more than 10 degrees in either direction. There was a response only as the slit moved across the screen from left to right and not in the reverse direction. Moreover, the response elicited from the left (ipsilateral) eye alone was weak.

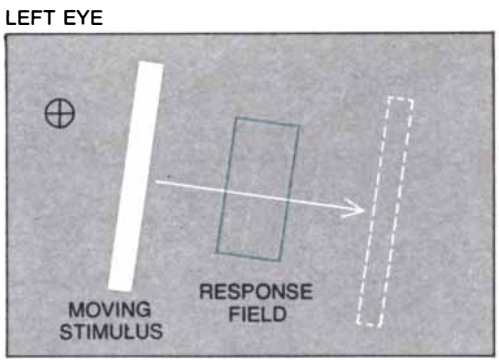
The relative strength of the responses elicited from each eye varies from one neuron to another, and the particular neuron shown in the illustration is a case of extreme contralateral dominance. The inhibitory contribution from the ipsilateral eye is far from weak, however, as can be seen by the reduction of the binocular response when the response fields are stimulated in the "wrong" spatiotemporal relationship, as for example when the prism setting is such that the response fields are side by side instead of being superimposed.

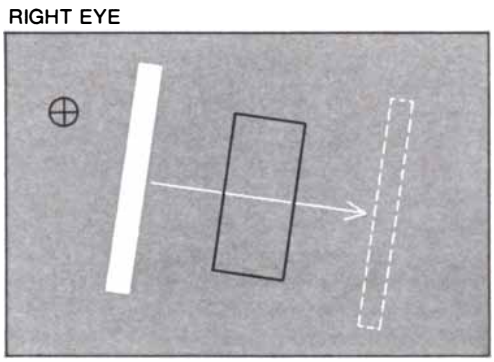
Our findings demonstrated that the binocular response of the neuron is in fact very sensitive to slight changes in the overlap of its response fields in the plane of the stimulus. Since the response fields themselves are quite small (less than half a degree), this means a high

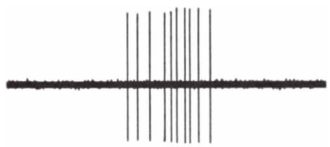


PENETRATION of the microelectrode through the cat's visual cortex is represented in this sequence of successively enlarged views. The top view of the entire brain (*a*) shows that the cortical region associated with vision (*color*) corresponds to a single gyrus, or fold, located toward the rear of each of the brain's two hemispheres. The cross section of this gyrus (*b*) reveals that the visual cortex is functionally divided into an array of tiny columns (*broken colored lines*) that run from the surface of the brain to the interior white matter. The magnified cross section (*c*) shows that each of these columns (which are not visible under the microscope) consists of a group of neurons with the same requirements for orientation of the stimulus. "Simple" neurons are those that appear to receive a fairly direct input from the fibers coming into the cortex. "Complex" neurons, which behave as if they receive an input from a large number of simple neurons, probably signal the output of the column.

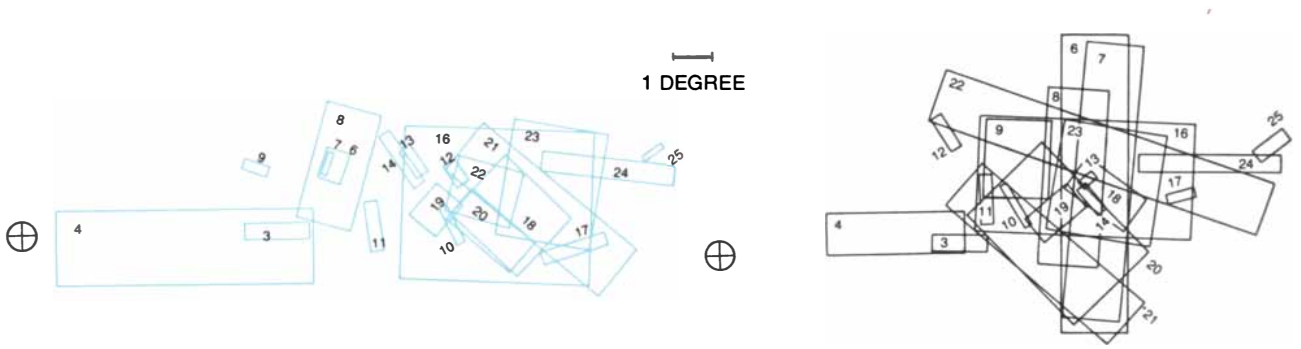








OSCILLOSCOPE TRACE



RESPONSE FIELDS FOR LEFT (IPSILATERAL) EYE

RESPONSE FIELDS FOR RIGHT (CONTRALATERAL) EYE

PAIRED RESPONSE FIELDS, one for each eye, can be plotted for a given neuron in the visual cortex of the cat (*top*). The response field of the neuron is defined as the region in the visual field of each eye where a specific stimulus will cause excitation of the cell (*middle*). In this illustration the response fields are separated on the screen because of the slightly divergent position the eyes assume in paralysis. Normally the fields would tend to overlap each other. By moving the microelectrode carefully through the cortex it is possible to record successively from a large number of different neurons; when the paired response fields of these neurons are plotted (*bottom*), those for the ipsilateral eye (the eye on the same side of the brain as the neuron in question) are more scattered than those for the contralateral eye (the eye on the opposite side). This means that it is not possible to superimpose all response fields on the same plane at the same time and that therefore different neurons would be optimally stimulated by objects in different planes.

level of disparity specificity. This particular neuron could indicate, by a marked decrease in firing rate, a disparity change as small as two minutes of arc, a feat approaching human performance. (The human threshold disparity is about 10 seconds of arc, or approximately 10 times better.)

It is perhaps not too surprising, in view of the very small size of the response fields, that the two retinal images of a binocularly presented stimulus must be very precisely located in order to produce a good response from the neuron. More surprising is the almost total suppression of the strong response from an appropriately located image in one eye if the image is inappropriately located in the other. This inhibition persists when the image is moved (for example by inserting the prism or by changing the distance of the stimulus) more than one degree of arc in either direction from the optimal position with respect to its correctly located partner in the other eye. In other words, binocular inhibition extends for more than one degree of retinal disparity on each side of the optimal disparity for a given neuron. The significance of this conclusion can be seen when one considers how nearby neurons behave with respect to one another in binocular vision; for those binocular neurons concerned with central vision the total range of optimal disparities is also a couple of degrees.

Let us now look at another disparity-specific binocular neuron recorded from the same column of tissue in the cortex as the one just described. Its stimulus requirements were quite similar (a slowly moving slit of light with the same orientation) except that the optimal disparity was 1.7 degrees more convergent because of the different position of its ipsilateral response field. Thus an oblique slit, in spite of the fact that it stimulates the contralateral response fields of both neurons, will under binocular viewing conditions excite one of them and inhibit the other, depending critically on its distance from the cat.

This binocular inhibition, operating over the same range as the range of disparity from one neuron to another, may be part of the explanation for the phenomenon of binocular fusion. A binocularly viewed target can be seen as being single in spite of the fact that it appears to lie in two different directions when the views from each eye alone are compared. In the upturned-bucket example, if the disparity between the retinal images of the larger circles is not too great, then one sees not two large circles but a single (fused) large circle floating in depth. It is reasonable to suppose that the failure to see a second large circle is due to the binocular inhibition of those neurons that were activated monocularly by such a circle. The narrowing down of the amount of activity among different neurons narrows down in turn the number of stimulus possibilities from which the brain has to choose. In this case groups of binocular neurons associated with the same contour but with different retinal disparities are narrowed to one group and therefore a particular disparity.

Both of the neurons described above belong to Hubel and Wiesel's class of simple neurons, that is, neurons that respond only to stimuli on narrowly defined areas of the retina. It was particularly interesting to examine the binocular properties of complex neurons, since they are thought to receive an input from a number of simple neurons and therefore to respond over a wider area of retina. Would they also respond over a wider range of disparity?

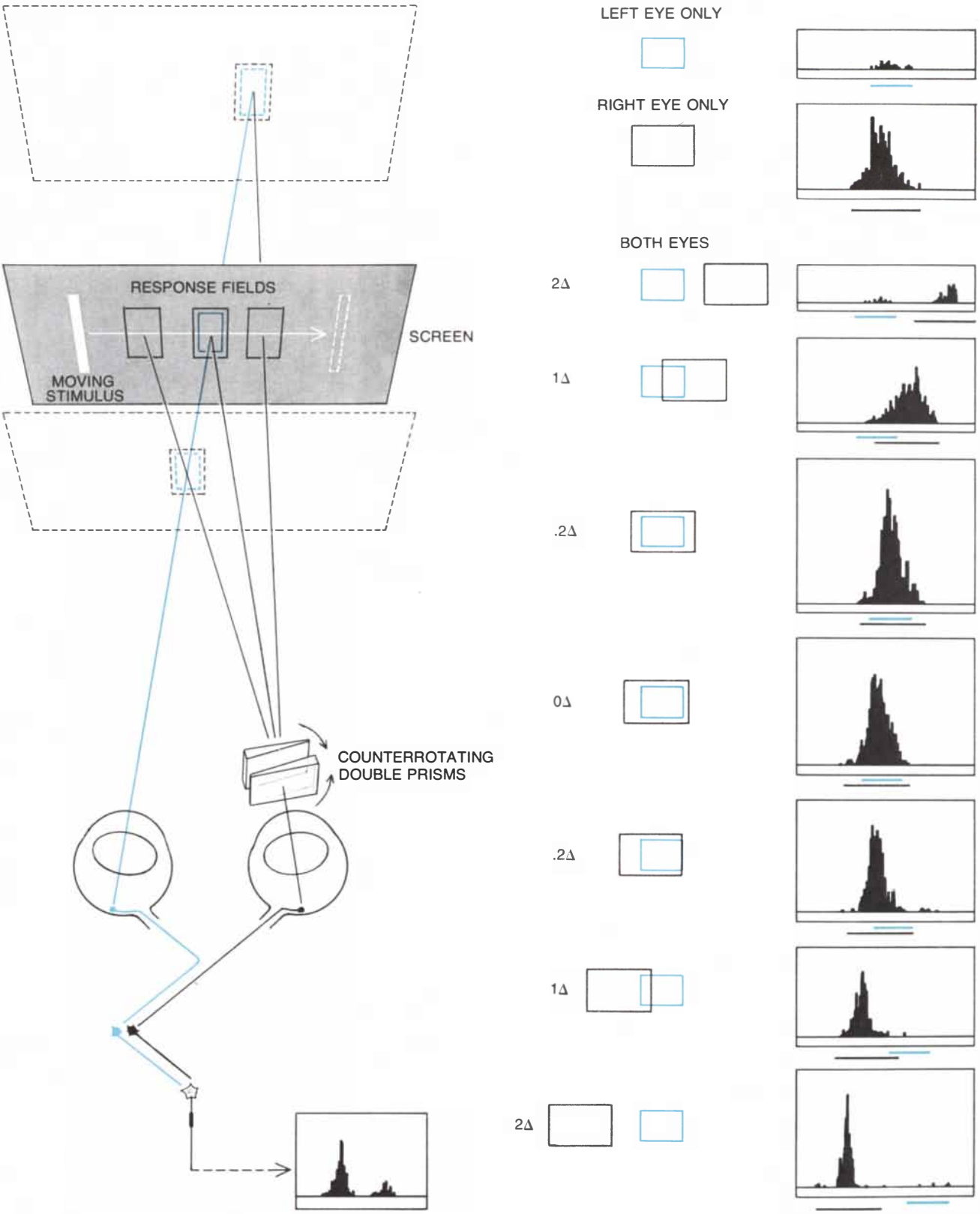
Two types of disparity-specific complex neuron were found. In one group there was a high degree of specificity in spite of the large size of the response field. One binocular complex neuron had response fields six degrees across but could still detect changes of disparity as accurately as most simple neurons (which have fields less than one degree across). This astonishing precision means that the neuron would signal with a change in firing rate that a stimulus moving anywhere over a six-degree area had produced a change of just a few minutes of arc in retinal disparity. With the eyes in a constant position a disparity-specific complex neuron "looks" at a thin sheet suspended in space and fires if a stimulus with the correct orientation and speed of movement appears anywhere on the sheet (but not in front of or behind it).

Disparity-specific complex neurons behave as if they receive an input from a number of simple neurons with different absolute response-field positions in each eye but with the same relative position, so that they all have the same optimal disparity. In fact, we noticed groups of such neurons in the Berkeley experiments, and Hubel and Wiesel have recently shown that binocular neurons with the same disparity specificity in the monkey's cortex appear to be grouped in cortical columns similar to the columns for orientation specificity. We therefore have another example of the cortical column as a system for extracting information about one specific type of stimulus while generalizing for others. A disparity-specific complex neuron can accordingly respond to a vertical edge moving over a wide region of the retina but over a very narrow depth in space. Directional specificity is lost but orientation and disparity information are retained.

There is some evidence for another type of binocular cortical column where all the neurons have response fields in the same position for the contralateral eye but have scattered fields for the ipsilateral eye. Blakemore calls these structures "constant direction" columns, because the neurons associated with them appear to respond at different disparities but to stimuli that are in the same direction from the contralateral eye. The output cell from such a column would presumably generalize for disparity but would be specific for the orientation and direction of the object.

Other complex binocular neurons responded over a wide range of disparity as well as of visual field. Since these neurons are active over the same range in which one observes binocular inhibition, they may be the source of the inhibition for simple neurons.

Once small residual eye movements had been accounted for and disparity specificity had been demonstrated, we were able to go ahead and compare the response-field pairs of a large number of different binocular neurons. In that way we could assess the total range of disparity variation. This was of particular interest because of a large body of observations obtained in psychophysical experiments on humans showing the range of disparity over which there is binocular fusion and the range of disparity over which stereopsis operates, both for central (foveal) vision and as one moves into the lower-resolution, peripheral visual field. The measurements were tedious because of the great length of time it takes to characterize a disparity-specific cortical neuron. In a typical experiment it took us three days to accumulate the 21 disparity-specific neurons whose response fields are shown in the illustration on the opposite page. All the neurons were recorded from the left striate cortex, and inspection of their response fields reveals a greater scatter in the fields of the left eye than in the fields of the right. This general observation that the ipsilateral receptive fields show more horizontal scatter is of interest in view of the fact that the ipsilateral



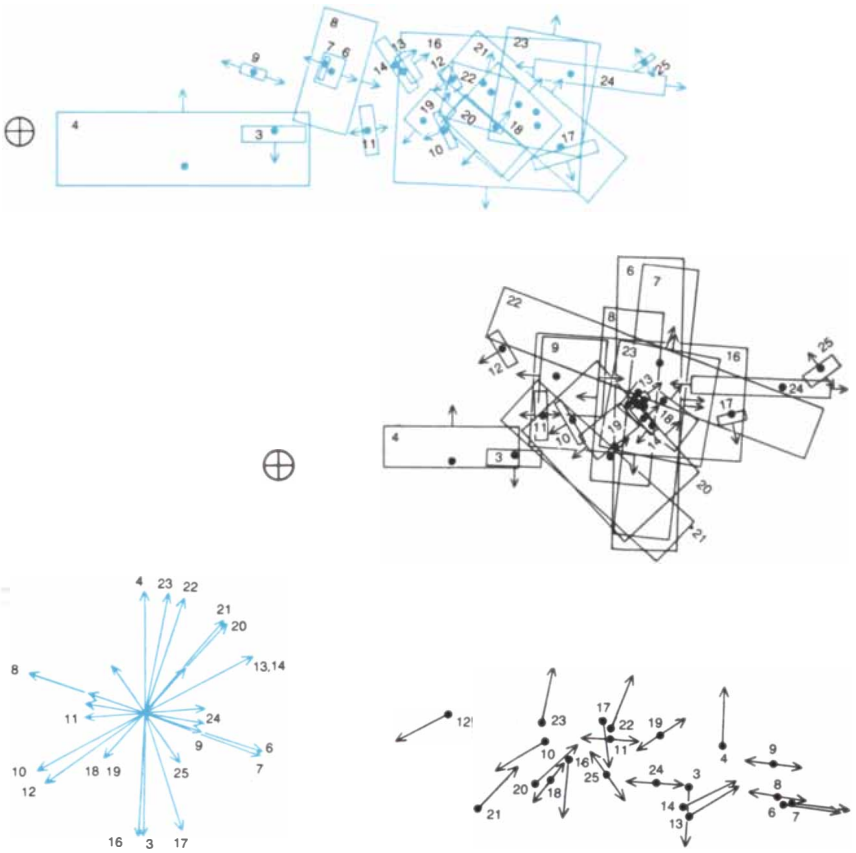
SPECIFICITY of a given binocular neuron for varying degrees of retinal disparity can be studied with the aid of counterrotating double prisms of variable power. The slight shifts of the two response fields with respect to each other in the plane of the moving stimulus are equivalent to setting that plane nearer to or farther from the animal (*left*). In this particular example, a case of extreme contralateral dominance, the response to a moving slit of light elicited from the left (ipsilateral) eye alone was weak; the inhibitory contribution from the ipsilateral eye is far from weak, however, as can be seen from the reduction of the binocular response when the prism setting is such that the response fields are side by side instead of being superimposed (*right*). The response for this particular neuron falls off rapidly for disparities on the order of tenths of a degree; hence there would be good discrimination among neurons whose optimal disparities cover a range of several degrees. A prism setting of one diopter (Δ) is equal to a retinal disparity of one centimeter at a distance of one meter; expressed as an angle, a disparity of one Δ equals .57 degree.

fibers are the ones that have arisen most recently in the evolution of binocular vision.

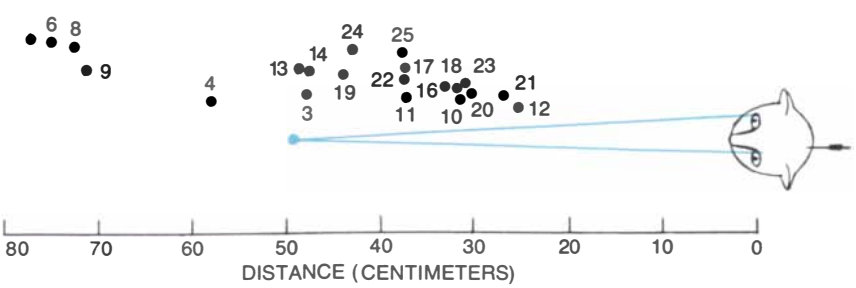
One can get a measure of the range of disparity in the response fields by shifting each pair of fields horizontally so that all the left-eye fields are superimposed. It is clear that if there were no disparity between different pairs of response fields, then that would lead to superimposition of the right-eye fields also. The degree to which the fields do not superimpose can be measured, and in this case there was a range of six degrees of horizontal disparity and two degrees of vertical disparity. It is not immediately obvious why the neurons should cover a range of vertical disparity, since only the horizontal component can be used for stereopsis. The psychophysical studies show, however, that although the visual system cannot make use of vertical disparity, allowance must be made for such disparities so that the system can still operate when they are introduced. Vertical disparities arise at close viewing distances (where the image of a given object may be significantly larger on one retina) and also in the course of eye movements (where the two eyes may not remain perfectly aligned vertically).

The total range of disparity surveyed by a given cortical area varies according to retinal eccentricity. Binocular neurons concerned with the area centralis (the high-resolution part of the cat's retina that corresponds to the human fovea) cover a disparity range of two degrees compared with six degrees for those neurons dealing with the visual field about 10 degrees away from the midline. The small total range for the area centralis not only allows fine discrimination within that range but also requires fine control of eye movements so that the target being examined can be kept within the range. The range of disparities for central vision appears to be even narrower in humans and monkeys, where there is exquisite control of convergent and divergent eye movements so that the midpoint of the range can be varied. The fineness of the range is attested by the double vision that occurs if there is the slightest imbalance in the muscular system.

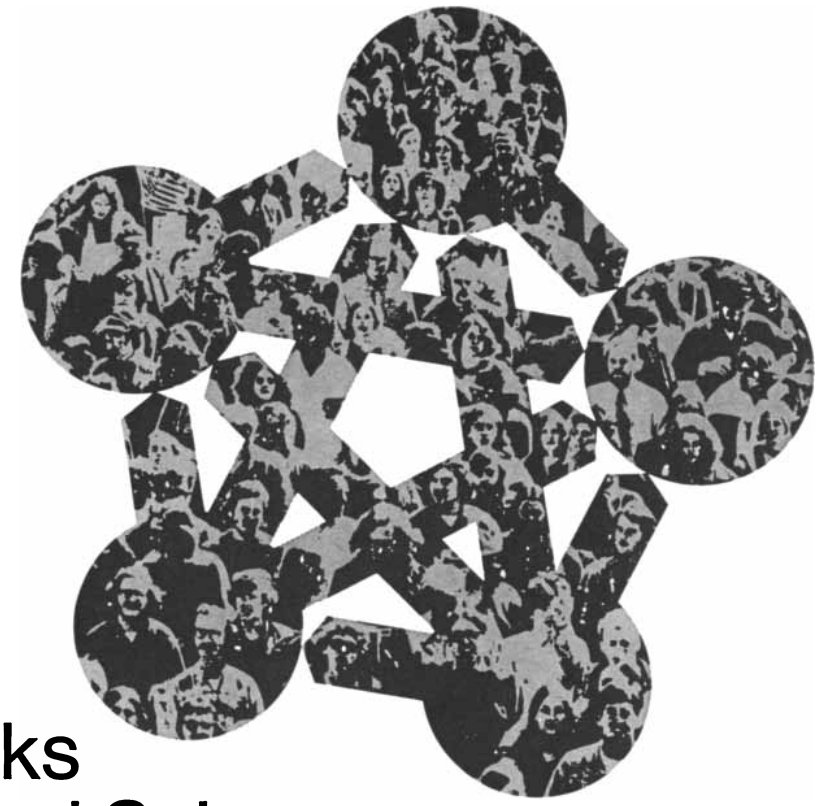
The preliminary results described here provide some insight into the initial operations performed by the visual cortex in extracting the information about disparity between small elements of the two retinal images. Much remains to be determined about how these first steps are utilized by the brain to yield our complete three-dimensional view of the world. Here are two examples of the kind of problem that remains to be solved: (1) Since convergent and divergent eye movements themselves produce changes in retinal disparity, how are these movements taken into account so that an absolute depth sense results that does not change with eye position? (2) How is a synthesis achieved from the disparity information about the myriad contours of a visual scene? The answers to these and many more perplexing questions about the brain may be best answered by the combination of the approaches of psychophysics and neurophysiology that has proved fruitful thus far.



RANGE OF OPTIMAL RETINAL DISPARITIES for the 21 binocular neurons whose response fields are plotted in the illustration on page 92 (and reproduced at the top of this illustration) can be calculated by shifting each pair of response fields horizontally so that the binocular centers (*black dots*) of all the left eye's fields are superimposed (*bottom*). The scatter of the right eye's response fields then gives the range of disparities. For those fields located away from the central area there is a range of six degrees of horizontal disparity and two degrees of vertical disparity. For neurons with fields closer to the center of the retina the range is smaller and hence the neurons are capable of finer discrimination.



DEPTH DISCRIMINATION attributable to binocular vision is represented by this plot of the points in space at which a correctly oriented contour would optimally stimulate the 21 binocular cortical neurons whose response fields are shown in the illustration at top of this page, provided that the cat's eyes are fixed on a point 50 centimeters in front (*colored dot*).



Important Books in the Behavioral Sciences:

ARMY ANTS

A Study in Social Organization

T. C. SCHNEIRLA, Late Curator,
The American Museum of Natural History
Edited by Howard R. Topoff,
The American Museum of Natural History

". . . Essential for anyone seriously studying the evolution of social behavior in insects. The army ants are perhaps the best example of group action and interdependence in insect colonies. Anyone publishing on army ants must be greatly indebted to and dependent upon the tremendous volume of data which Schneirla assembled in his lifetime. This book is a guide to those data, including some heretofore unpublished observations and a summary of numerous papers. It is a prime source of what is known and a stockpile of ideas for future research." —*Science,* January 28, 1972

1971, 349 pages, 89 illustrations including 8 full color plates, $12.00

INTRODUCTION TO MULTIVARIATE ANALYSIS FOR THE SOCIAL SCIENCES

JOHN P. VAN DE GEER, University of Leiden, The Netherlands

Without entering into the more complex study of statistical tests, this text provides an introduction to multivariate techniques with emphasis on their interrelations and on basic understanding of the mathematical procedure. It is designed to teach seniors and graduate students in the social and behavioral sciences the principles of multivariate analysis relevant to their fields. Recommended especially for advanced courses in psychology, sociology, industrial relations, and education.

1971, 293 pages, 57 illustrations, $9.50

THE BIOLOGICAL AND SOCIAL MEANING OF RACE

RICHARD H. OSBORNE, University of Wisconsin

In terms understandable to the layman, 13 distinguished scientists discuss the current state of knowledge about race. While emphasizing social factors, they treat this complex subject within the context of population concepts and the science of genetics.

1971, 182 pages, illustrated, cloth $6.00, paper $2.95

DEVELOPMENT AND EVOLUTION OF BEHAVIOR

Essays in Memory of T. C. SCHNEIRLA

Edited by LESTER R. ARONSON and ETHEL TOBACH
The American Museum of Natural History, and
DANIEL S. LEHRMAN and JAY S. ROSENBLATT,
Institute of Animal Behavior, Rutgers University

"A tribute to a genuine pioneer in the interdisciplinary field between psychology and biology; to a man whose importance in this role has long been underrated except by those who were fortunate enough to be his pupils. . . .

"This volume testifies about his influence behind the scenes, and makes clear how many lines of research he has pursued and inspired. Edited by four of his most prominent pupils, it contains no fewer than 28 contributions by an extremely wide range of scientists, young and old, and almost all of these contributions are more than simple reports on isolated studies; every one of them is thought-provoking—Schneirla made people think hard."
—N. Tinbergen in *Nature,* November 28, 1970

1970, 656 pages, 30 illustrations, $12.95

From your bookseller, or from

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104
58 Kings Road, Reading, England RG1 3AA

Cyclic AMP

This comparatively small molecule is a "second messenger" between a hormone and its effects within the cell. It operates in cells as diverse as bacteria and cancerous animal cells

by Ira Pastan

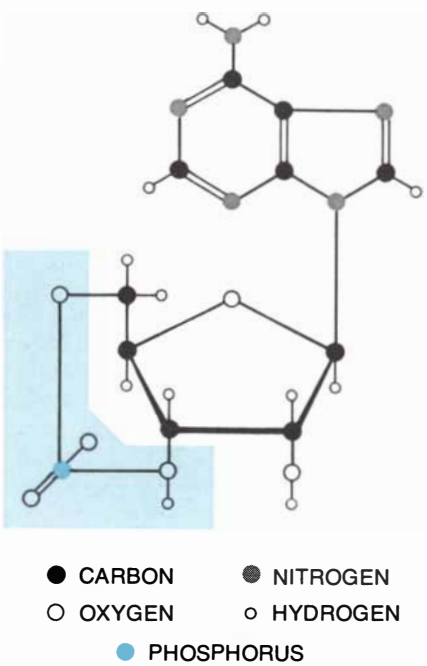
The chemical reactions that proceed in the living cell are catalyzed by the large molecules called enzymes. If all the enzymes found within cells were working at top speed, the result would be chaos, and many mechanisms have evolved that control the speed at which these enzymes function. A small molecule that plays a key role in regulating the speed of chemical processes in organisms as distantly related as bacteria and man is cyclic-3′5′-adenosine monophosphate, more widely known as cyclic AMP. ("Cyclic" refers to the fact that the atoms in the single phosphate group of the molecule are arranged in a ring.) Among the many functions served by cyclic AMP in man and other animals is acting as a chemical messenger that regulates the enzymatic reactions within cells that store sugars and fats. Cyclic AMP has also been shown to control the activity of genes. Moreover, a precondition for one of the kinds of uncontrolled cell growth we call cancer appears to be an inadequate supply of cyclic AMP.

The first steps leading to the discovery of cyclic AMP were taken by Earl W. Sutherland, Jr., at Washington University some 25 years ago. For this work Sutherland was awarded the Nobel prize in physiology and medicine for 1971. Sutherland was trying to trace the sequence of events in a well-known physiological reaction whereby the hormone epinephrine (as adrenalin is generally known in the U.S.) causes an increase in the amount of glucose, or blood sugar, in the circulatory system. It is this reaction, usually a response to pain, anger or fear, that provides an animal with the energy either to fight or to flee. It is not a simple one-step process. Glycogen, a polymeric storage form of glucose, is held in reserve in the cells of the liver. What transforms the glycogen into glucose, which can then leave the liver and enter the bloodstream, is a series of steps involving intermediate substances. Sutherland measured the levels of these intermediates and concluded that only the initial step in the series (the transformation of glycogen into the intermediate sugar glucose-1-phosphate) was mediated by epinephrine. The transformation itself is actually catalyzed by an enzyme known as phosphorylase. Observing the activity of phosphorylase in cell-free extracts of liver tissue, Sutherland was able to enhance the enzyme's performance by first exposing to epinephrine the cells he used to make his extract.

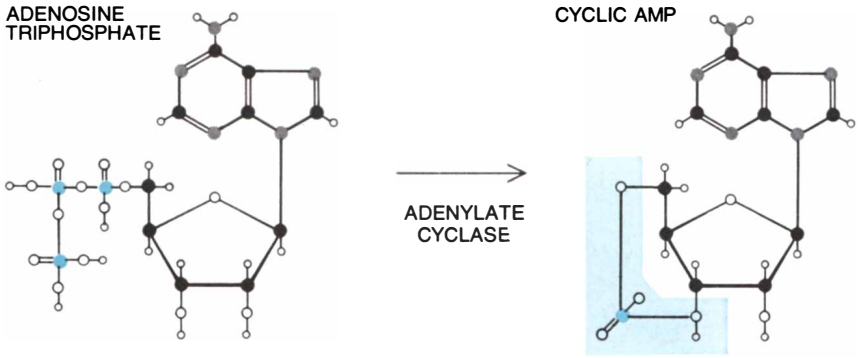
Sutherland began to examine the properties of phosphorylase in more detail. He found that the enzyme could exist in two forms: one that degraded glycogen rapidly and one that had no effect on it. The conversion of the enzyme from the active to the inactive form was catalyzed by a second enzyme, whose only action was to remove inorganic phosphate from the phosphorylase molecule. The conversion is worth noting; it is an important example of how the activity of an enzyme can be controlled by a relatively small change in its structure.

In collaboration with the first of a number of talented co-workers, Walter D. Wosilait, Sutherland next found still another liver enzyme, phosphorylase *b* kinase, which could restore the inactive form of phosphorylase to the active state. As one might expect, the reversal was accomplished by replacement of the missing phosphate; the donor of the phosphate was a close chemical relative of cyclic AMP, adenosine triphosphate (ATP). At about the same time Edwin G. Krebs and Edmond H. Fischer of the University of Washington detected a similar activating kinase in muscle tissue. ("Kinase" is the name reserved for transformations where ATP is the phosphate-donor.)

Having established the existence of two forms of phosphorylase, Sutherland and his colleagues concluded that the speed at which glycogen was broken down in the liver was a function of the amount of the enzyme present in its active form. Sutherland and Theodore W. Rall now made preparations of ruptured liver cells. When they added epinephrine to these broken cells, they found that in spite of the damage the hormone still increased the activity of the enzyme. The experiment, although simple, was extremely significant. Never before had a hormone been observed to



CYCLIC AMP is so named because the phosphate group in its molecule (*colored area*) forms a ring with the carbon atoms to which it is attached. Earl W. Sutherland, Jr., and his colleagues isolated substance in 1958.



FORMATION OF CYCLIC AMP takes place when an enzyme in the cell membrane, adenylate cyclase, responds to the arrival of a hormone at the membrane. The enzyme transforms molecules of adenosine triphosphate, or ATP (*left*), within the cell into cyclic AMP (*right*).

function in a preparation that contained no intact cells. Once such a reaction can be shown to occur in a cell-free preparation the investigator can go on to test various cell components one at a time to determine just which ones are affected by the hormone.

This Sutherland and his co-workers proceeded to do. They knew that the phosphorylase was present in liver-cell cytoplasm: that part of the cell outside the nucleus and inside the cell membrane. When they added epinephrine to preparations composed of cytoplasm alone, however, there was no increase in enzyme activity. The absence of response suggested that the hormone exerted its effect on some other component of the liver cell. In due course they found that this component was the cell membrane.

Exactly what was the hormone doing to the cell membrane? In an effort to find out Sutherland employed a stratagem commonly used in biochemistry. He incubated a preparation of cell membrane (which itself contains no phosphorylase) with epinephrine. He then brought the mixture to the boiling point, expecting the heat to destroy the activity of the enzyme or enzymes in the membrane that were required for epinephrine action. When he added the now denatured mixture to a cell-free preparation that contained phosphorylase but no cell membrane, the activity of the phosphorylase was increased. The epinephrine had interacted with the cell membrane during the initial incubation period, evidently causing some enzyme in the membrane to produce a heat-stable factor that enhanced the activity of phosphorylase. Unfortunately the factor—whatever it was—was present in very small amounts and was therefore difficult to identify.

Sutherland eventually collected a large enough sample of the factor to determine that it belonged to the group of small molecules known as nucleotides. It did not, however, appear to be any of the known nucleotides. He wrote to Leon A. Heppel of the National Institutes of Health, who had developed many of the methods used in preparing and identifying a number of nucleotides, asking him for a quantity of an enzyme that breaks nucleotides into their component parts and thus facilitates their identification. This request set the stage for a remarkable coincidence.

It is Heppel's habit to let letters that do not require an immediate answer accumulate on his desk. He covers each few days' correspondence with a fresh sheet of wrapping paper, and every few months he clears his desk. Heppel immediately sent Sutherland the enzyme but left the letter of request on his desk. By chance the stratum just below contained a chatty letter from a friend and former colleague, David Lipkin of Washington University, describing an experiment where ATP was treated with a solution of barium hydroxide. The result was the formation of an unusual nucleotide.

Heppel remembers coming into his office one Saturday to clear his desk. He found Sutherland's and Lipkin's letters in adjacent strata and consequently reread them together. It seemed likely to him that both men had isolated the same substance, and he proceeded to put them in touch. Lipkin's chemical synthesis readily produced the nucleotide in large quantities. This made it easy to establish that the synthetic substance was structurally identical with natural cyclic AMP. It also provided an abundant supply of synthetic cyclic AMP for experimental purposes.

Taking advantage of the demonstrated ability of cyclic AMP to increase phosphorylase activity in cell-free preparations, Sutherland and his co-workers were able to measure the amount of cyclic AMP present in a wide variety of cells. They found that it was 1,000 times less abundant than ATP, being present in cell water in a ratio of about one part per million in contrast to ATP's one part per 1,000. Although scanty in amount, cyclic AMP was present in virtually every organism they examined, from bacteria and brine-shrimp eggs to man. Among mammals it was present in almost every type of body cell. Sutherland's group went on to examine a number of tissues that are characterized by their secretion of various substances following stimulation by hormones. He discovered that the level of cyclic AMP in such cells rose soon after exposure to the hormone. Moreover, when the tissues were exposed to nothing but cyclic AMP, or to derivatives of cyclic AMP that enter cells rapidly, they produced secretions just as readily.

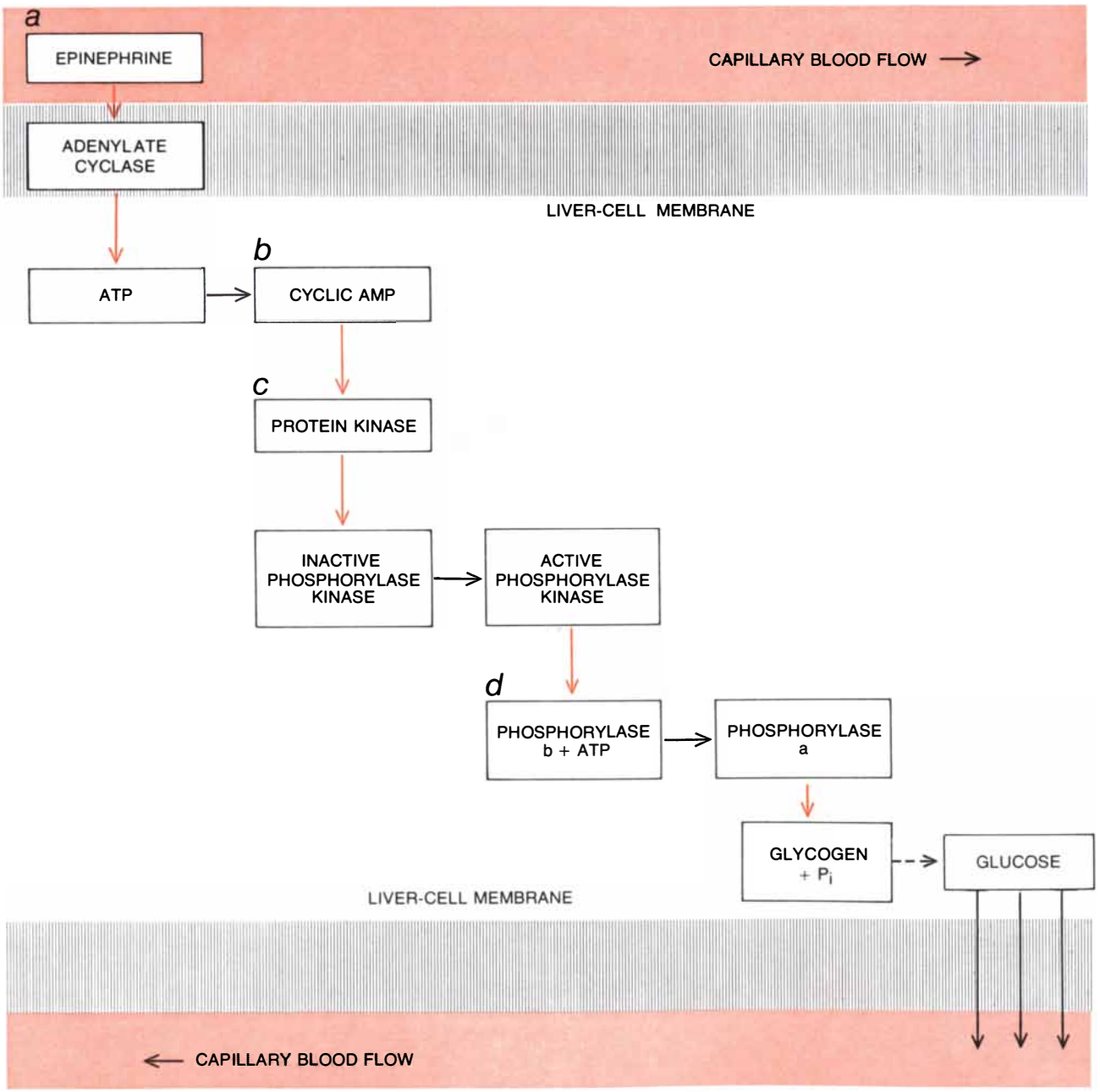
The cyclic AMP molecule is formed from ATP by the action of a special enzyme: adenylate cyclase. The enzyme is located in the membrane of the cell wall. Normally its activity is low and the transformation of ATP into cyclic AMP takes place at a slow rate. Let us consider what happens, however, when a hormone enters the bloodstream. The hormone acts as a "first messenger." It travels to the target cell and then binds to specific receptor sites on the outside of the cell wall. Thyroid cells have receptors that "recognize" thyrotrophin, adrenal cells have receptors that recognize adrenocorticotrophin, and so forth. The binding of the hormone to the receptor site increases the activity of the adenylate cyclase in the cell membrane; just how this occurs has not yet been established. In any event, cyclic AMP is produced as a result, utilizing the abundant supply of ATP on the inner side of the cell membrane. The cyclic AMP is then free to diffuse throughout the cell, where it acts as a "second messenger," instructing the cell to respond in a characteristic way. For example, a thyroid cell responds to this second message by secreting more thyroxine, whereas an adrenal cell responds by producing and secreting steroid hormones. In the cells of the liver the instruction results in the conversion of glycogen into glucose.

Because cyclic AMP is such a powerful regulator of cell functions the cell must be able to control its level of concentration. In most cells control is accomplished by regulating the rate of synthesis of cyclic AMP and by the actions of one or more enzymes, known as phosphodiesterases, that degrade cyclic

AMP into an inert form of adenosine monophosphate. The deactivation results from a splitting of the ester bond that joins the phosphate to the 3′ carbon of the deoxyribose ring. The quantity of the degradative enzymes in the cell is not kept constant; apparently more can be made whenever the level of cyclic AMP in the cell is elevated for more than a few minutes. The level of cyclic AMP is also controlled by diffusion through the cell wall; this is the mechanism operating in addition to enzyme degradation in bacteria and in the cells of some animal tissues.

Krebs, whose earlier work with Fischer had established the presence of active and inactive forms of phosphorylase in muscle tissue, was able in 1968 to specify the role played by cyclic AMP in activating the enzyme. Working with Donal Walsh, he found that the cyclic AMP binds to yet another enzyme, protein kinase, which is inactive until cyclic AMP is present. The activated kinase then performs the same function for a related enzyme, phosphorylase kinase. It is this second enzyme that at last activates the phosphorylase. The result of the final activation is the breakdown of glycogen, the storage form of glucose, in a series of steps similar to those that proceed in liver cells.

Whenever glycogen is being degraded in order to satisfy the organism's need for glucose, it would be a waste of energy to continue the synthesis of additional glycogen. This waste is avoided. A specific enzyme mediates the syn-



RELEASE OF BLOOD SUGAR by a glycogen storage cell in the liver is mediated by cyclic AMP. In the schematic diagram arrows in color symbolize actions and black arrows the results. "First messenger," epinephrine, arrives at the cell membrane and activates the enzyme, adenylate cyclase (*a*), causing it to convert some of the ATP present in the cytoplasm into cyclic AMP (*b*), the second messenger. The cyclic AMP then activates a protein kinase (*c*), which activates a second kinase. The second kinase (*d*) triggers a four-step sequence (*not shown*) that converts the glycogen into the assimilable sugar glucose, which then passes into the bloodstream.

thesis of glycogen. It is glycogen synthetase, and like phosphorylase and many other enzymes it has an active and an inactive form. At the time when some molecules of cyclic AMP are initiating the chain of events that leads to the breakdown of glycogen, other cyclic AMP molecules are at work converting glycogen synthetase from the active to the inactive form.

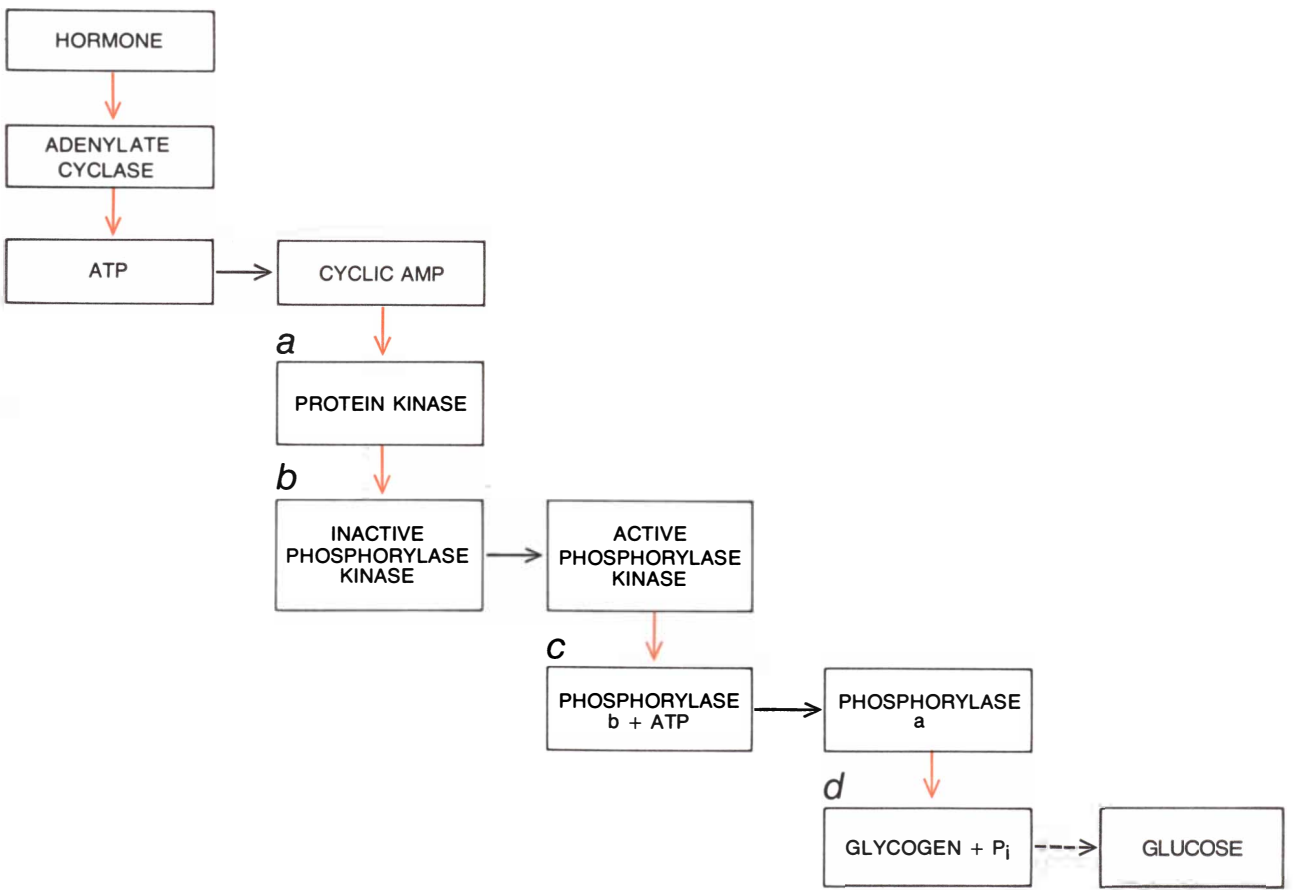
Just as cells that store glycogen have the task of supplying the fasting organism with glucose, so the task of fat cells is to satisfy the organism's need for fatty acids. Fatty acids are present in the fat cell in the storage form triglyceride. The triglyceride can ocupy as much as 90 percent of the cell's volume. Here again it is cyclic AMP that initiates the breakdown of the stored fat. In response to any of several hormonal stimuli the level of cyclic AMP in the fat cell begins to rise. As with muscle tissue, this activates a protein kinase. The kinase in turn activates a second enzyme, triglyceride lipase. On being converted to the active form the second enzyme begins to degrade the stored triglyceride into the required fatty acids. It should be noted here that protein kinases are present in the cells of many tissues other than fat and muscle; numerous other actions of cyclic AMP presumably also involve these ATP-powered enzymes.

In addition to such effects within cells cyclic AMP has been observed to stimulate the expression of genetic information. How this stimulation is accomplished in animal cells remains obscure. Almost all the detailed observations of the regulation of gene activity involve a single microorganism: the common intestinal bacterium *Escherichia coli*. There are many good reasons for molecular biologists who are engaged in genetic studies choosing to work with this simple organism. One of them is that a typical animal cell contains enough DNA to account for 10 million individual genes; *E. coli* has only enough DNA for about 10,000 genes.

Sutherland and another colleague, Richard Makman, established the presence of cyclic AMP in *E. coli* cells in 1965. By then the substance was already known to control a variety of cell processes, and Robert Perlman and I at the National Institutes of Health guessed that it must also play an important role in the bacterium. But what role?

There were two clues. First, cultures of *E. coli* that are nourished exclusively with glucose show low levels of cyclic AMP. Second, such cultures can synthesize only very small amounts of a number of enzymes, including those needed to metabolize sugars other than glucose; this inhibition is called the "glucose effect." Putting these clues together, we reasoned that cyclic AMP was a chemical switch, so to speak; if it was present in *E. coli* in adequate quantities, it would activate the expression of those genes that are necessary for the synthesis of the missing enzymes. In order to test this speculation we added cyclic AMP to cultures of inhibited *E. coli*. The cells were then able to metabolize such sugars as maltose, lactose and arabinose in addition to a variety of other nutrients.

Now, factors that operate at the level of the gene do so by stimulating the synthesis of the messenger RNA that in



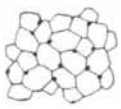
INITIAL SEQUENCE that triggers the transformation of glycogen into glucose in muscle tissue also involves cyclic AMP. First (*a*) the cyclic AMP that has been formed from ATP in the cell cytoplasm activates a protein kinase. The kinase in turn activates a second kinase (*b*) that is capable of transforming phosphorylase *b* into phosphorylase *a*. When this takes place (*c*), the transformed phosphorylase then starts the sequence (*d*) that converts glycogen into glucose. Edwin Krebs and Edmond Fischer of the University of Washington discovered the phosphorylase alteration in muscle at the same time that Sutherland found the same process in the liver.

effect reproduces the information contained in the DNA of the gene. The messenger RNA then provides the ribosomes of the cell with the information they need to construct the appropriate enzyme. In our laboratory by this time we had isolated a mutant form of *E. coli* that lacked the cell-membrane enzyme required for the synthesis of cyclic AMP. Curiously, the mutant cells were viable; apparently the presence of cyclic AMP is not absolutely crucial to survival. For our next experiment we selected cultures of the mutant strain that contained a known amount of the two messenger RNA's needed for the metabolism of two sugars: lactose and galactose. When we added cyclic AMP to the mutant cultures, we found that the quantity of the two messenger RNA's was increased but that the expression of most of the other genes in the cells was not affected. In order to learn exactly how the increase took place we now needed to study the reaction in a cell-free preparation.

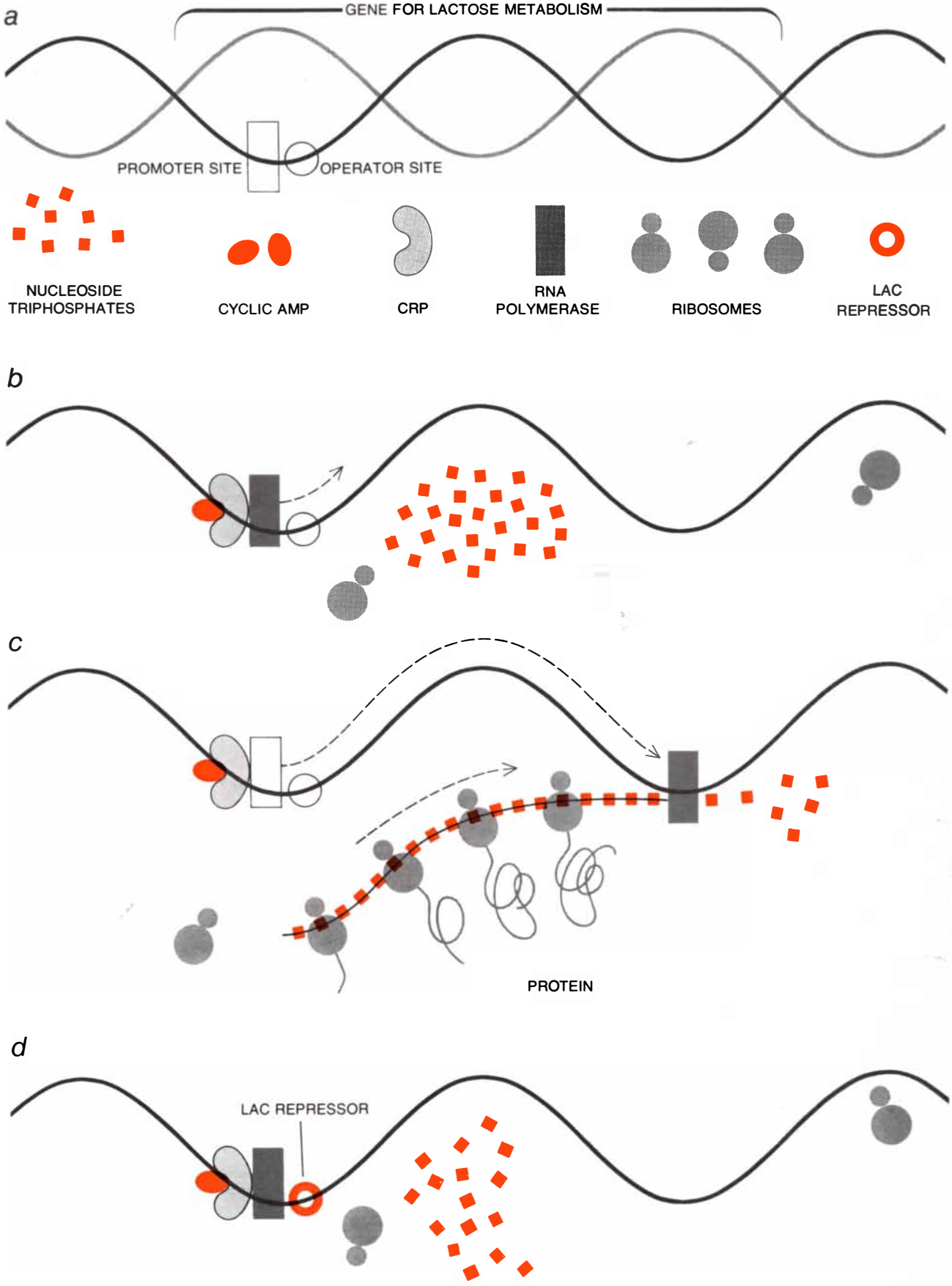
At Columbia University, Geoffrey L. Zubay and his co-workers were working with a complex cell-free preparation made from *E. coli*. When they added to the preparation DNA that was greatly enriched in the genes for lactose metabolism, small amounts of one of the enzymes of lactose metabolism, beta-galactosidase, were formed. The DNA was enriched because it was derived from a bacterial virus that had acquired the lactose genes from the *E. coli* host it had once lived in and now contained it permanently. The virus is a hybrid of two other viruses, designated λ and 80, and the DNA derived from it is $\lambda 80$*lac* (for lactose) DNA.

When Zubay and his co-worker Donald Chambers added cyclic AMP to the cell-free preparation, synthesis of the enzyme for lactose metabolism was greatly stimulated. Soon thereafter my colleagues B. de Crombrugghe and H. Varmus showed that the synthesis of *lac* messenger RNA was also increased. John Parks in our laboratory, following Zubay's procedure, developed a cell-free *E. coli* preparation that responded to the addition of λ*gal* (for galactose) DNA by producing one of the enzymes of galactose metabolism. Like Zubay's preparation, Parks's synthesized much more of the enzyme when cyclic AMP was added.

One might now have expected that the same result could be achieved without even using the complex cell-free preparation. It would be necessary only to mix together in the test tube appropriate quantities of DNA rich in lactose genes and of the special enzyme that

TISSUE	HORMONE	PRINCIPAL RESPONSE
FROG SKIN 	MELANOCYTE-STIMULATING HORMONE	DARKENING
BONE 	PARATHYROID HORMONE	CALCIUM RESORPTION
MUSCLE 	EPINEPHRINE	GLYCOGENOLYSIS
FAT 	EPINEPHRINE	LIPOLYSIS
	ADRENOCORTICO-TROPHIC HORMONE	LIPOLYSIS
	GLUCAGON	LIPOLYSIS
BRAIN 	NOREPINEPHRINE	DISCHARGE OF PURKINJE CELLS
THYROID 	THYROID-STIMULATING HORMONE	THYROXIN SECRETION
HEART 	EPINEPHRINE	INCREASED CONTRACTILITY
LIVER 	EPINEPHRINE	GLYCOGENOLYSIS
KIDNEY 	PARATHYROID HORMONE	PHOSPHATE EXCRETION
	VASOPRESSIN	WATER REABSORPTION
ADRENAL 	ADRENOCORTICO-TROPHIC HORMONE	HYDROCORTISONE SECRETION
OVARY 	LUTEINIZING HORMONE	PROGESTERONE SECRETION

FOURTEEN EXAMPLES of hormonal activities that affect many different target tissues (*left*) have one factor in common: each causes an increase in the level of cyclic AMP in the tissue. It seems probable that all the responses are set in train by the sudden increase in level.



copies DNA into RNA, add some ATP and other nucleoside triphosphates as building blocks and cyclic AMP as mediator and the result would be the production of *lac* messenger RNA. The expectation proved to be false. Some RNA was formed in the test tube but none of it corresponded to the *lac* gene. Clearly some factor was missing. Fortunately there was a clue to its nature.

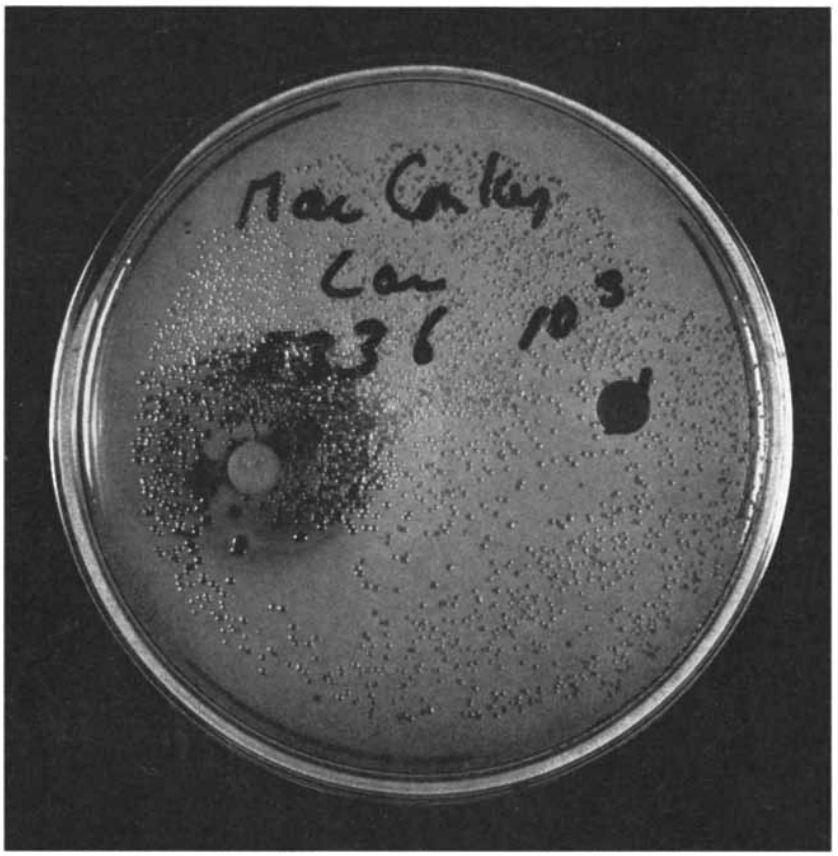
In our search for mutant strains of *E. coli* that could not make various enzymes known to be controlled by cyclic AMP we had found some mutants that produced cyclic AMP in abundance but were still unable to metabolize either lactose or several other sugars. Similar *E. coli* mutants had been isolated at the Harvard Medical School by Jonathan Beckwith and his colleagues. What was lacking in the mutants was a protein that has the ability to bind cyclic AMP. The protein, which is known either as catabolite gene activation protein (abbreviated CAP) or as cyclic AMP receptor protein (CRP), was difficult to purify but a pure form was finally prepared by my colleague Wayne B. Anderson.

De Crombrugghe added some of the pure protein to the test-tube mixture that had failed to yield *lac* messenger RNA. This time the effort was successful. Moreover, when CRP was added to a similar test-tube mixture that included λ*gal* DNA, *gal* messenger RNA was formed. These experiments, incidentally, were the first to achieve the transcription of a bacterial gene in a system containing only purified components. We could now proceed to examine in detail how the gene activity was controlled.

The initial step in the process involved the combination of cyclic AMP with CRP. The complex of the two substances was then bound to the DNA; once this took place the enzyme that copies DNA into RNA was enabled to bind to a specific site, called the "promoter," at the beginning of the *lac* gene in the bacterial DNA. After that the nucleoside triphosphates initiated the *lac* transcription process. We soon found that the transcription of the genes for galactose metabolism followed the same steps.

Of the 10,000 or so *E. coli* genes, perhaps some hundreds are regulated in this way. How the others are controlled remains a mystery. The regulation of even those genes that are controlled by cyclic AMP involves other substances as well. The *lac* transcription process provides an example. A specific protein, *lac* repressor, is bound to the DNA at a site (called the operator) at the beginning of the *lac* gene. The operator site is near the site on the DNA where the complex of cyclic AMP and CRP binds. The cell cannot now produce *lac* messenger RNA even in the presence of CRP and cyclic AMP because the repressor prevents the copying enzyme from beginning the transcription [*see illustration on opposite page*]. Only when the repressor is removed by the action of a substance closely related to lactose can the transcription take place.

This pattern of events, where cyclic AMP is able to stimulate the transcription of many genes while at the same time individual repressors can prevent certain transcriptions, gives the *E. coli* cell the flexibility it needs for the efficient utilization of the foodstuffs in its environment. *E. coli* does not store large amounts of carbohydrate and must live on what is in its immediate vicinity. If the cell is exposed, say, both to glucose (which it can already metabolize) and to



MUTANT STRAIN of the bacterium *Escherichia coli* lacks the enzyme that turns ATP into cyclic AMP. As the different reactions show, addition of cyclic AMP to one disk in the dish (*left*) enabled the colonies of bacteria nearby to metabolize certain sugars. The bacteria near the second disk (*right*), to which only 5′ AMP was added, were not able to do this.

TRANSCRIPTION OF A GENE, the sequence of events shown schematically on the opposite page, occurred in a cell-free medium that was supplied (*a*) with RNA polymerase (*solid bar*), nucleoside triphosphates (*colored squares*), the protein CRP (*crescent shape*), which reversibly binds cyclic AMP (*colored ovals*), and quantities of DNA enriched with the gene for lactose metabolism (*bracketed area of helix*). Transcription begins when a combined unit of CRP and cyclic AMP activates a promoter site at the beginning of the *lac* operon (*b*); RNA polymerase now binds to the promoter site, ready to link the nucleoside triphosphates together in the correct sequence. As transcription proceeds (*c*), the RNA polymerase arranges the nucleoside triphosphates in the correct sequence. Ribosomes meanwhile begin the process of assembling the proteins that comprise the enzymes for lactose metabolism. If still another protein, *lac* repressor (*colored doughnut*), attaches itself to the operator site at the beginning of the *lac* operon (*d*), then the RNA polymerase can bind but cannot transcribe in spite of the presence of the CRP-cyclic AMP unit. This pattern, whereby cyclic AMP stimulates the transcription of many different operons, whereas the repressors are specific for a particular gene, makes for flexibility in the utilization of foodstuffs.

lactose (which it cannot), there is nothing to be gained by expending energy to produce the enzymes for lactose metabolism. As long as its supply of cyclic AMP remains at a low level the *E. coli* cell conserves its capacity for protein synthesis.

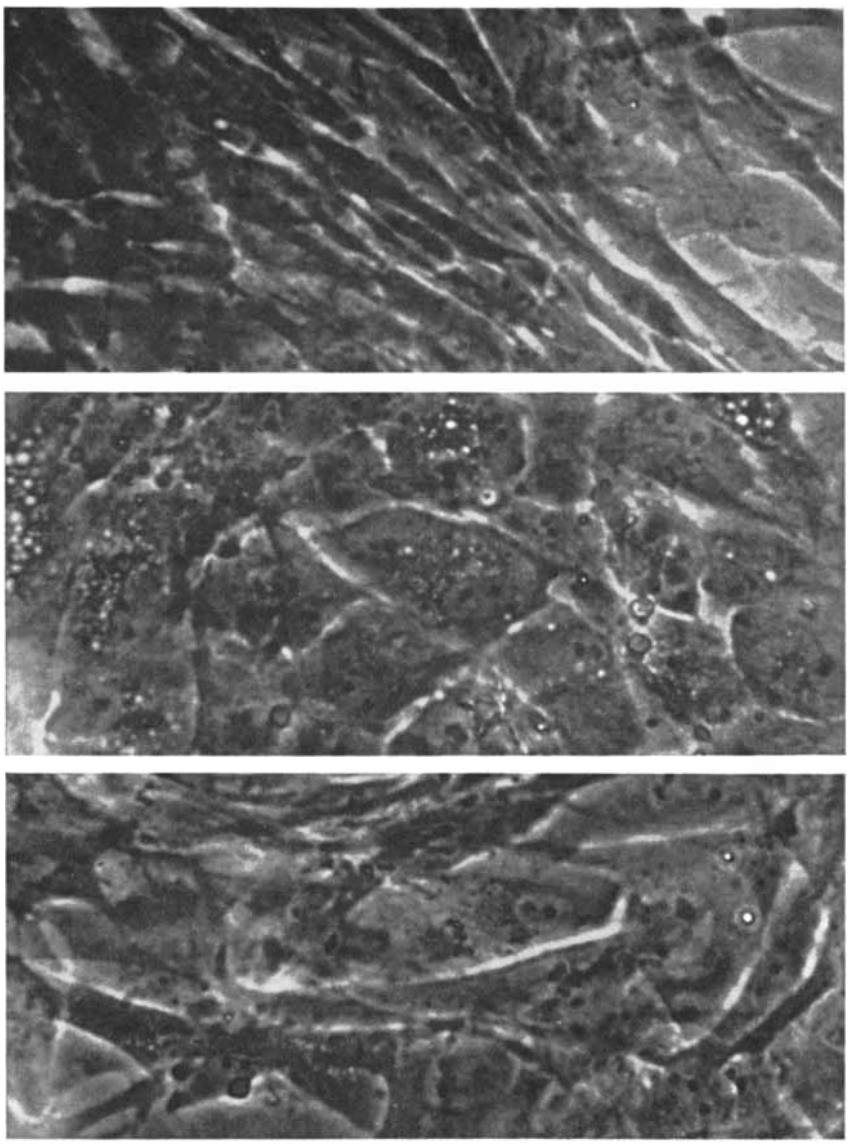
That so many different cell functions should be under the control of a single substance is remarkable. Cyclic AMP also participates in the process of visual excitation and regulates the aggregation of certain social amoebae so that they can form complex reproductive structures [see "Hormones in Social Amoebae and Mammals," by John Tyler Bonner; SCIENTIFIC AMERICAN, June, 1969]. A good example of the substance's versatility is provided by the contrast between the mechanisms of glycogen and lipid degradation on the one hand and the stimulation of gene transcription on the other. The degradation reactions involve enzymes and depend on protein phosphorylation. In *E. coli* gene transcription no phosphorylation is involved. It will be interesting to learn whether the *E. coli* mechanism is employed to control gene activity elsewhere in nature and whether cyclic AMP acts in still other ways. Most of the substance's actions in animal cells remain to be explored.

Abnormalities in the metabolism of cyclic AMP may explain the nature of certain diseases. For example, in cholera the bacteria responsible for the disease produce a toxin that stimulates the intestinal cells to secrete huge amounts of salt and water; it is the resulting dehydration, if it is unchecked, that is fatal [see "Cholera," by Norbert Hirschhorn and William B. Greenough III; SCIENTIFIC AMERICAN, August, 1971]. It appears that the toxin first stimulates the intestinal cells to accumulate excess cyclic AMP, whereupon the cyclic AMP instructs the cells to secrete the salty fluid.

In our own laboratory we are interested in understanding the difference between normal cells and cancer cells. Since cancer cells typically grow in an uncontrolled manner and lose their ability to carry out specialized functions, it seemed possible to us that some of their abnormal properties might be attributable to their inability to accumulate normal amounts of cyclic AMP. My colleague George Johnson and I have begun to investigate this possibility.

The cells that are commonly used for such studies are fibroblasts: cells that contribute to the formation of connective tissue. They are usually taken from the embryos of chickens, mice or other animals and allowed to grow in a nutrient fluid. After a short period of culture in a medium approximating blood serum in composition, the embryonic cells take on the appearance of the normal fibroblast cells in connective tissue and grow in bottles or dishes in a controlled manner. When the cultured cells are exposed to cancer-producing viruses or to chemical carcinogens, however, they begin to grow in an abnormal manner. They take on the appearance of tumor cells, and when they are injected into a suitable host, they usually produce tumors. The process of changing a normal cell to a cancer cell in culture is termed transformation. Among the properties of transformed cells are a change in appearance, accelerated growth, looser adherence to the container surface, alterations in the rate of production of specialized large molecules such as mucopolysaccharides and clumping on exposure to certain agglutinative plant proteins.

Now, cells that have been transformed and are then grown in the presence of cyclic AMP tend to return to normal. They grow more slowly, adhere more tenaciously to the container, synthesize certain large molecules at a faster rate and clump less when they are exposed to agglutinative plant protein. Their appearance also frequently returns toward normal. As far as we now know, the morphologic reversal occurs only in embryo cells and those derived from con-



CHICK EMBRYO CELLS infected with a temperature-sensitive mutant strain of Rous sarcoma virus maintain a normal appearance (*top micrograph*) when they are cultured at a temperature of 40.5 degrees Celsius. When the temperature is reduced to 36 degrees, however, they quickly develop an abnormal appearance (*middle*). If cyclic AMP is added to the cells, their appearance remains normal (*bottom*) even after the temperature is reduced.

nective tissue; tumors from a few other kinds of cell do not show the same morphologic response.

In any event the reversal suggested to us that transformed cells might contain abnormally low levels of cyclic AMP. One of our co-workers, Jack Otten, investigated this possibility. He found that cells from chick embryos that had been transformed by exposure to Rous sarcoma virus contained much less cyclic AMP than normal chick-embryo cells. We could not be certain, however, which came first: the abnormal appearance of the transformed cells or the low level of cyclic AMP.

To settle the question we needed a way to transform cells very rapidly. We had at our disposal a mutant variety of the Rous sarcoma virus that made rapid cell transformation possible. For example, chick cells infected with the mutant virus (which had been isolated by our colleague John Bader of the National Cancer Institute) remain normal in appearance as long as the culture is kept at a temperature of 40.5 degrees Celsius. When the temperature is reduced to 36 degrees, however, the cells are rapidly transformed.

We grew cultures infected with this temperature-sensitive virus and kept them at the "normal" temperature. We incubated some of the cultures with a potent derivative of cyclic AMP and some without it. When we lowered the temperature of the cells incubated with cyclic AMP to the transformation level, they continued to look normal for some time [*see illustration on opposite page*]. The cultures without cyclic AMP, once the temperature was lowered, developed the characteristic transformed appearance within a few hours.

Otten measured the level of natural cyclic AMP in the readily transformed cells after their exposure to the lower temperature. He found that as soon as 20 minutes later the level had fallen greatly; this was well before the cells began to develop a transformed appearance.

This finding obviously leaves many questions unanswered. Is the phenomenon confined to tumors of connective tissue? How many of the abnormal properties of the transformed cells result from the low level of cyclic AMP? Which enzymes are responsible for lowering the level? Are there other tumor-forming viruses and chemical carcinogens that similarly lower the cell's supply of cyclic AMP? Not least, can the findings to date be exploited in a therapeutically useful manner? The search for answers continues in each of these areas.

MATHEMATICAL GAMES

The curious properties of the Gray code and how it can be used to solve puzzles

by Martin Gardner

The binary Gray code is fun,
For in it strange things can be done.
Fifteen, as you know,
Is one, oh, oh, oh,
And ten is one, one, one and one.
—ANON.

Although the decimal system is now in common use throughout the world, mathematicians and computers often manipulate integers by using other systems, some with such exotic features as mixed bases, negative bases, irrational bases or floating points. One of the most useful of these systems—one with surprising puzzle applications—is the Gray code.

The first puzzle application of a Gray code, which I shall describe below, was in 1872, when a binary version provided an elegant solution to a much older mechanical puzzle. (The origin of the binary version and its application to telegraphy are discussed in the article by F. G. Heath beginning on page 76.) The term Gray, however, derives from Frank Gray, a research physicist at the Bell Telephone Laboratories, who died in 1969. His contributions to modern communication technology were immense. The method now in use for compatible color television broadcasting was developed by Gray (numerologists note!) in the 1930's. In the 1940's he devised what was soon to be called the binary Gray code to avoid the large errors that could arise in transmitting signals by pulse code modulation (PCM). The first publication of this code was in his U.S. Patent 2632058 (March 17, 1953) for a Gray coder tube that eliminated the quantizing grid wires used in early PCM transmission tubes.

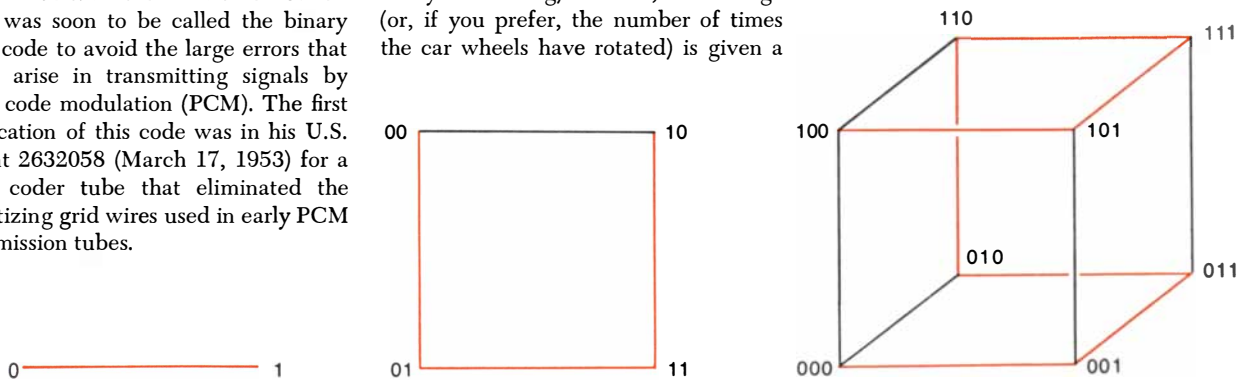
Exactly what is a Gray code? It is a way of symbolizing the counting numbers in a positional notation so that when the numbers are in counting order, any adjacent pair will differ in their digits at one position only, and the absolute difference at that position will be 1. For instance, 193 and 183 could be adjacent counting numbers in a decimal Gray code (the middle digits differ by 1), but not 193 and 173, nor 134 and 143. There is an infinity of Gray codes, since they apply to any base system and for each base there are many different ways to construct the code.

To appreciate the value of such a system, consider what happens when the odometer of a car reads 9,999 miles. To register the next mile, five wheels must rotate to show 10,000. Because the wheels move slowly, there is little chance of error. But if counting is recorded electronically at enormously high speeds, when two or more digits change simultaneously the likelihood of producing a false number zooms upward. The probability is greatly reduced if the counting procedure requires only one decision whenever the magnitude to be coded is halfway between two adjacent quantized steps, regardless of whether the magnitude is increasing or decreasing. If the counting is by Gray code, only one digit of the counter changes by only one unit at each step.

The mileage meter is a familiar example of what are called analogue-to-digital converters. A continuous (in this case always increasing) variable, the mileage (or, if you prefer, the number of times the car wheels have rotated) is given a digital output. There are many other control systems in which analogue-to-digital conversion must proceed at enormously high speed while the variable being measured fluctuates rapidly. Examples include wind-tunnel simulations of airplanes and guided missiles, and PCM applications where voltages, shaft positions, wave amplitudes of sounds, colors and so on must be translated almost instantly to a digital output signal. At one time a human observer would take pointer readings or inspect a curve on a graph, record the magnitude in digital form and feed this information to a computer. Today the slow and error-prone middleman is eliminated by analogue-to-digital converters connected directly to the computer. A great increase in accuracy and often a considerable saving in hardware result from counting scales in Gray codes.

Binary Gray codes are the simplest. If we limit the code to one digit, there are only $2^1 = 2$ numbers, 0 and 1. Disregarding reversals, there is only one Gray code: 0, 1. We can graph this as a straight line, its ends labeled 0 and 1 [*see illustration at left below*]. The Gray code is obtained by moving along the line in either direction. A Gray code for two binary digits has $2^2 = 4$ numbers: 00, 01, 10 and 11. The corners of a square can be labeled with these numbers [*see middle illustration below*]. The labeling is such that the binary numbers at any pair of adjacent corners differ in only one place. We can start at any corner and visit all four corners by going clockwise or counterclockwise around the square. Ignoring reversals, this produces four Gray codes. The colored line starting at 00 yields the Gray code 00, 01, 11, 10. The code is cyclic because the path can return to 00.

A Gray code for three-digit binary numbers has $2^3 = 8$ numbers that can be placed on the corners of a cube [*see illustration at right below*]. Adjacent cor-



Graphs for two-digit (left), *three-digit* (middle) *and four-digit* (right) *binary Gray codes*

ners have binary triplets that differ in only one place. Any continuous path that visits every corner once only generates a Gray code. For example, the path shown by the colored line starting at 000 produces 000, 001, 011, 010, 110, 111, 101, 100. This is a cyclic code because the path can return from 100 to 000 in one step. Such paths are called Hamiltonian paths after the Irish mathematician William Rowan Hamilton. As the reader has probably guessed, binary Gray codes correspond to Hamiltonian paths on cubes of n dimensions. A Gray code for four-digit binary numbers has $2^4 = 16$ numbers that fit the corners of a hypercube in 4-space, for five digits a hypercube in 5-space, and so on. Interested readers will find this covered in detail in E. N. Gilbert's "Gray Codes and Paths on the n-Cube," in the *Bell System Technical Journal* (Volume 37, Number 1, pages 815–826, May, 1958).

Gray codes for other bases correspond to Hamiltonian paths on more complicated n-dimensional graphs. The number of Gray codes for any base increases explosively as the number of digits increases. The number of Gray codes, even for the binary system, is known only for four or fewer digits.

A recent ill-fated attempt to obtain the number for five binary digits is recounted by Ronald C. Read in *Graph Theory and Its Applications,* edited by Bernard Harris (Academic Press, 1970). Read had written a BFI program for finding the number of Hamiltonian paths on the five-dimensional cube. BFI is Read's acronym for brute force and ignorance. ("It should be BFBI," he has since remarked, "the second B standing for 'Bloody,' but one has to preserve a measure of decorum in published papers.") "These are algorithms," he explains, "devoid of any subtlety whatever, which simply keep thumping the problem on the back until it disgorges an answer." After running the program for a short time (on a computer in Kingston, Jamaica) a sample of the output was examined in order to estimate how long the run would be. The guess was 10 hours, and so the computer was set to run overnight unattended. During the night a tropical thunderstorm cut the power supply and the computer stopped.

"Idle curiosity," Read continues, "prompted us to look to see where the program had got to before being so abruptly terminated, and in doing so we discovered that we had made a rather serious error in calculating our previous estimate of the running time. Our revised estimate turned out to be more like ten years!"

Read sensibly abandoned the project and we still do not know in how many ways a five-dimensional fly could walk along the edges of a five-dimensional cube, visiting each of its 32 corners once only, thereby generating all the binary Gray codes for five digits.

For practical purposes it is important to select a Gray code with two desiderata: (1) rules for its formation should apply to the entire set of counting numbers; (2) it should have simple conversion rules for translating a standard number to its Gray code equivalent and vice versa.

The simplest Gray code with both features is called a reflected Gray code. For most mathematicians it is *the* Gray code. To convert a standard binary number to its reflected Gray equivalent start with the digit at the right and consider each digit in turn. If the next digit to the left is even (0), let the former digit stand. If the next digit to the left is odd (1), change the former digit. (The digit at the extreme left is assumed to have a 0 on its left and therefore remains unchanged.) For example, applying this procedure to binary number 110111 gives the Gray number 101100.

To convert back again consider each digit in turn starting at the right. If the sum of all digits to the left is even, let the digit stay as it is. If the sum is odd, change the digit. Applying this procedure to 101100 restores the original binary number 110111.

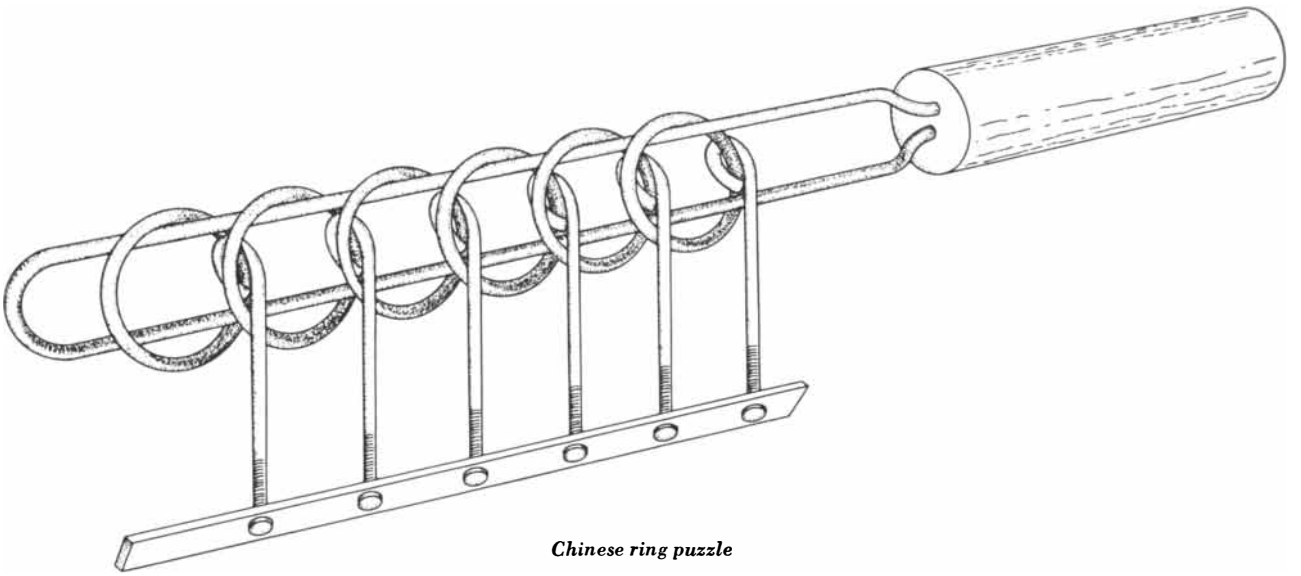
Inspection of the numbers from 0 through 42 and their reflected binary Gray code equivalents will show that every two adjacent Gray numbers differ at only one place, and of course the difference is necessarily 1 [*see illustration at right*]. It is called a reflected code because the series can be generated rapidly by the following algorithm. Start with 0, 1 as a one-digit Gray code, then reflect (reverse) and append the digits to get 0, 1, 1, 0. Next put 0's in front of the first two numbers and 1's in front of the last two numbers. The result is a two-digit Gray code: 00, 01, 11, 10. To extend the series to three-digit Gray numbers, reflect the two-digit code: 00, 01, 11, 10, 10, 11, 01, 00. As before, put 0's in front of the first half of these numbers and 1's in front of the last half: 000, 001, 011, 010, 110, 111, 101, 100. This corresponds to a Hamiltonian path starting at 000 on a cube.

Proceeding in this fashion, first reflecting the entire series, then adding 0's and 1's on the left, one can quickly generate the reflected binary Gray code to any desired counting number. Note that for each set of n-tuplets the code is cyclic in that the first and last n-tuplets also differ at only one spot. If the code is used by a counter consisting of wheels, such as the usual mileage meter, the meter can go from its highest number back to zeros with a final unit change of only one wheel.

In 1872 Louis Gros published in Lyon a brochure on *Théorie du Baguenodier.* "Baguenodier" (more commonly spelled

	FEDCBA
0	0
1	1
2	11
3	10
4	110
5	111
6	101
7	100
8	1100
9	1101
10	1111
11	1110
12	1010
13	1011
14	1001
15	1000
16	11000
17	11001
18	11011
19	11010
20	11110
21	11111
22	11101
23	11100
24	10100
25	10101
26	10111
27	10110
28	10010
29	10011
30	10001
31	10000
32	110000
33	110001
34	110011
35	110010
36	110110
37	110111
38	110101
39	110100
40	111100
41	111101
42	111111

Reflected binary Gray code for 0 through 42



Chinese ring puzzle

"baguenaudier") is the French name for a classic puzzle known in the English-speaking world as Chinese rings, although any connection between the puzzle and China is unknown to me. In his brochure Gros applied a binary notation to this puzzle for the first time. The puzzle had been first described in 1550 by Girolamo Cardano in his *De Subtilitate Rerum,* and it was later analyzed at considerable length by John Wallis in his *Algebra* in 1693.

Several versions of the Chinese rings (the number of rings can vary) are currently on sale in the U.S. One sturdy six-ring aluminum version about 11 inches long can be obtained for $4.95 postpaid from Watson Products, Box 636, Altadena, Calif. 91001. If you are handy with tools, the puzzle can be made with curtain rings, stiff wire and a strip of wood with holes drilled through it [*see illustration above*].

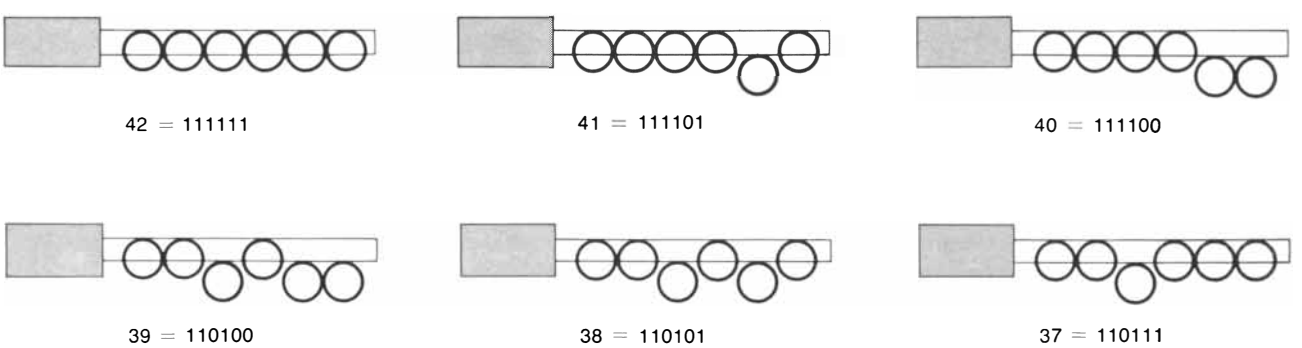
The object of the puzzle is to free all the rings from the double bar. As a first move the two end rings can be dropped either individually or both at once. To simplify the solution we shall assume that only one of the two end rings is dropped at a time. With the exception of those two rings (which can always be taken off or put on simultaneously), a ring will come off or go on only when its immediate neighbor closer to the end is on and all the other rings beyond are off. This is the peculiar feature of the puzzle that makes it so frustrating and repetitious.

Let each ring be represented by a binary digit, 1 for on, 0 for off. The binary Gray number for 42 [*see illustration on preceding page*] is 111111. If we let this represent the six rings on the upper rod, each Gray number going from 42 back to 0 shows which ring is to be removed or put on to solve the puzzle in a minimum number of moves! It is apparent that for n rings, to determine the number of moves required, we simply write n as a Gray number of n units, convert it to standard binary and so obtain the answer. In this case the Gray number 111111 corresponds to 101010 in standard binary, which is 42 in decimal notation. (Gros explained all this in a slightly different way, but it amounts to the same thing.) To find the number by formula, use $\frac{1}{3}(2^{n+1} - 2)$ when n is even and $\frac{1}{3}(2^{n+1} - 1)$ when n is odd.

We have assumed that for each move only one ring is removed or put on. The brackets in the illustration on the preceding page indicate pairs of moves that can be made simultaneously with the two end rings. If these are counted as single moves, the six-ring puzzle can be solved in 31 moves instead of 42. The formulas for this "fast way" of solving an n-ring puzzle are $2^{n-1} - 1$ if n is even and 2^{n-1} if n is odd.

With a six-ring puzzle the slow-to-fast ratio is $42/31 = 1.355$; for seven rings it is $85/64 = 1.328$. The ratios continue 1.338, 1.332, 1.334, ... N. S. Mendelsohn has shown that this oscillating series converges rapidly on $1\frac{1}{3}$. Twenty-five rings require 22,369,621 steps. Assuming that a skilled operator can do 50 steps a minute, he could solve the puzzle the slow way, working 10 hours per day, in a little more than two years. By doing it the fast way, however, he could cut the time by about half a year.

As far as I know, Jesse R. Watson,



First six positions for solving ring puzzle using the Gray code

who sells the ring puzzle, was the first to ask and answer the following question: Suppose the initial position for an n-ring puzzle has the last ring (the one nearest the handle) on and all other rings off. Watson calls this the position of "maximum effort" because it requires more moves than any other position to take all the rings off. Assuming that the slow method is used, what simple formula gives the required minimum number of moves? (The answer will be given next month.)

The binary Gray code also solves the well-known Tower of Hanoi puzzle, in which n disks of diminishing sizes are stacked in a pyramid. The problem is to transfer them one at a time to a second spot, using a third spot as a temporary resting place with the proviso that no disk be placed on top of a smaller disk. (See Chapter 6 of *The Scientific American Book of Mathematical Puzzles & Diversions.*) To solve this puzzle for five disks, label the disks of the initial pyramid, starting with the smallest, from A to E. Label the columns of the illustration on page 107 from A to F as shown. Take the Gray numbers in sequence. At each step move the disk that corresponds to the column in which there is a change of digit. The sequence begins *ABACABAD*.... On every move a disk can be transferred to only one spot. The sequence solves the puzzle in $2^n - 1$ moves, which in this case is 31.

Rules for converting numbers in other base systems to reflected Gray numbers are simple generalizations of the rules for binary numbers. (There are several general conversion procedures but I give the simplest here.) If the base is even, the rules are the same as for the binary system except that when a digit is altered, it is changed to its "complement" with respect to $n - 1$ when n is the base, that is, to its difference from $n - 1$. In the binary system, n minus 1 equals 1, so that this means a simple change of 0 to 1 or 1 to 0. In the decimal system, numbers are complemented with respect to 9 (that is, subtracted from 9). Therefore to convert a decimal number to a Gray number take each digit in turn beginning at the right. If the next digit to the left is even, leave the former digit unchanged. If the left digit is odd, complement the former digit. For example, 1972 becomes 1027. To convert back to the decimal system, work with sums. If all digits to the left have an even sum, let the digit stand. If the sum is odd, subtract the digit from 9.

Only a slight modification of rules is required for numeral systems with an odd base. In such cases the sum rule applies to conversion in either direction. In the ternary system, for instance, complementation is with respect to 2. Regardless of which way you convert, complement when the sum on the left is odd, otherwise let the digit stand. Ternary Gray numbers, in counting order, are 0, 1, 2, 12, 11, 10, 20, 21, 22, 122, 121, 120, ...

In all bases, Gray counting numbers of the reflecting type (unless otherwise specified, these are considered *the* Gray numbers for a given base) are quickly determined by generalizing the procedure given for binary numbers. This is best explained by using the decimal system as an example [*see illustration at right*]. Note that the unit's column begins with the sequence 0 through 9, then it proceeds from 9 through 0, then from 0 through 9 and so on. In the 10's column, 10 0's (not shown) are followed by 10 1's, then by 10 2's, 10 3's and so on through 10 9's until 99 is reached. Now the doublets are reflected after every 100 steps, and in the third column from the right 100 0's are followed by 100 1's, then by 100 2's and so on until 999 is reached. The reader should have little difficulty applying this procedure to other base systems. In the ternary system, for example, reflections occur in the right column every third step, in the next column every ninth step, in the next column every 27th step and so on through increasing powers of 3.

Because Gray codes are relatively unknown to students of recreational mathematics, I suspect they have many puzzle applications other than the ones given here. I would be glad to hear from readers who know of recreational uses for Gray codes with bases greater than 2.

	GRAY
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	19
11	18
12	17
13	16
14	15
15	14
16	13
17	12
18	11
19	10
20	20
21	21
22	22
23	23
24	24
25	25
26	26
27	27
28	28
29	29
30	39

Reflected decimal Gray code

Daniel S. Fisher, a high school student in Ithaca, N.Y., generalized David L. Silverman's generalization of the betting problem given in April. He showed that Silverman's game is fair for any division of ownership of the coins when values of the coins are 1, n, n^2,..., n^k and the coins are weighted to fall tails with probability $1/n$ (n equal to or greater than 2).

So many readers sent other preposterous explanations of parallel ski tracks going around a tree—the last of April's problems—that I can give only a sampling:

A small, supple tree that bent as the skier went over it was proposed by John Ferguson, John Ritter, Brad Schaefer, Oliver G. Selfridge and James Weaver. Ferguson also suggested (among his 23 possibilities) a pair of skis pulled uphill by long ropes and two toboggan teams of *very* small midgets, four on each **ski.** Selfridge included this one: The **skier,** aware of his ineptness, wore a protective lead suit. His impact on the tree sheared out a cylindrical section. The dazed skier passed between top and bottom parts of the tree before the top fell down and balanced perfectly on the base.

Manfred R. Schroeder, director of the Drittes Physikalisches Institut at the University of Göttingen, reported an actual experience he had in 1955 while skiing down a mountain in New Hampshire. "I hit a small but sturdy tree with my right shinbone. The binding came loose and the ski and leg went around different sides of the tree. Below the tree, leg and ski came together again. However, the binding did not engage (no automatic step-in bindings then!) and the tracks ended in a spill about ten yards farther down the slope. Even then, in spite of considerable pain in the leg, I thought it was a worthwhile experience."

THE AMATEUR SCIENTIST

A compact short-focus telescope with spherical optical surfaces

Conducted by C. L. Stong

Telescopes scarcely larger than a shoe box that are capable of high resolution and magnification in excess of 100 diameters have become increasingly popular in recent years. The compact design is achieved by folding incoming rays of light with an optical system that consists of a pair of mirrors and a correcting lens. Few amateurs have attempted to make such telescopes, chiefly because the optical surfaces of the more popular designs must be ground and polished in the form of deep paraboloids or ellipsoids—figures that tax the skills of experienced optical workers.

This difficulty is circumvented in a telescope developed recently by Robert J. Magee of Concord, Mass. With the mathematical technique of ray tracing Magee designed a set of spherical surfaces that accomplish the same objective. Although it is not necessarily easy to grind and polish glass to a spherical figure, the surfaces can be made by an inexperienced worker who is patient and persistent. Magee describes the construction as follows:

"My objective in developing this design was to achieve an optical system that would be physically short in relation to its effective focal length. I also wanted the resolution to be limited primarily by the wave nature of the light. As can be judged from the accompanying diagram [*opposite page*] the physical distance between the primary mirror and the secondary mirror is about 1.66 times the diameter of the aperture. When the thickness of the primary mirror and the space required for a diagonal mirror and an eyepiece are taken into account, the overall length of the system is roughly twice the diameter of the aperture.

"The optical system consists of a perforated primary mirror, a two-element corrector lens (with one surface aluminized to function as the secondary mirror), a diagonal mirror and the eyepiece. The small two-element corrector lens replaces the full-aperture lens of the popular Maksutov system. The combination of an achromatic lens and a second-surface mirror that I use is known as a Mangin mirror.

"Assume that light enters the instrument from the left. Rays reflected by the primary mirror proceed through the corrector lens and fall on the secondary mirror. After being reflected from this surface the rays return through the corrector lens and come to a focus about five inches to the right of the primary mirror. A front-surface mirror or a prism can be inserted immediately behind the perforation of the primary mirror to divert the rays at a right angle into the eyepiece.

"The dimensions listed in the accompanying table [*page 112*] are scaled for a primary mirror 4¼ inches in diameter. The resolution of the system will remain diffraction-limited if all dimensions are altered in proportion as the aperture of the primary mirror is increased, although the aberration known as coma will severely limit the useful field of view at an aperture of eight inches.

"It might seem that the five optical surfaces of this small telescope are both more costly and more difficult to grind and polish than the single surface of a larger instrument of the Newtonian type, which is traditionally made by amateurs. The economy of the system results from the small size of the glass blanks and from the modest cost of the mounting. In terms of performance the instrument is comparable to a Newtonian telescope of the same aperture and is far more convenient to transport and use.

"The order in which the various surfaces are ground and polished can be varied according to the worker's preference. The sequence I followed is not necessarily the best one. I shall describe the operations in that order, however, so that I can point out the pitfalls on the basis of firsthand experience.

"The scale of optical systems can in general be altered within reasonable limits without sensibly impairing performance: the radii of curvature, the thickness of elements and the spacing between elements can be changed from the calculated values if all are kept in proportion. With this requirement in mind I ground the primary mirror first and scaled the remainder of the system accordingly. Later I learned by experience that an error of plus or minus .2 inch in the radius of curvature of the primary mirror need not be taken into account by altering other dimensions.

"An error of this size does change the back focal length of the optical system and the optimum distance between the primary mirror and the corrector-lens assembly. An increase in the radius of the primary mirror increases the back focal length. One can compensate for such an error by adjusting the distance between the primary mirror and the corrector lens while assembling the elements to the mounting.

"The central hole in the primary mirror was partially cut in the rear of the blank before the reflecting surface was ground. The minimum diameter of the perforation is about .9 inch. Optically the primary mirror is relatively fast. The focal ratio is $f/2.2$. The radius of the mirror was monitored frequently during the grinding operation by a center-of-curvature test. The test apparatus consisted of a point source of light formed by an illuminated pinhole in a piece of white cardboard. The primary mirror was positioned so that the pinhole occupied a point near the center of curvature of the mirror. When the mirror was wet, rays from the pinhole were reflected by the mirror and converged toward the cardboard.

"The cardboard is moved toward or away from the mirror as necessary to focus a sharp image of the pinhole on the cardboard. The mirror is turned as required to move the image close to the pinhole. The distance from the surface of the mirror to the image is measured. It equals the radius of curvature of the mirror.

"The measurements are not particu-

larly accurate during the early stages of grinding because rough glass is not a good optical surface even when wet. The image of the pinhole appears as a fuzzy spot, but it becomes increasingly sharp as the grinding progresses through successively finer grades of abrasive. A sharply focused image can be observed at any stage of the rough grinding by slightly polishing the glass. Make up a conventional pitch lap coated with optical rouge or cerium oxide. Place the mirror on top of the lap and push it back and forth about 25 times in each direction. Small flat areas will be polished at the tips of peaks that form the roughly ground surface. Collectively the polished areas will reflect diverging rays from the pinhole with sufficient sharpness for measuring the radius to within .05 inch. The finely ground surface is polished to a spherical figure and tested with the conventional techniques employed by amateur telescope makers.

"I next made the corrector-lens assembly, beginning with surfaces R_3 and R_4. The radii of these surfaces are equal. The surfaces are cemented together after the lenses are polished. Lens 2, which faces the incoming light, is used as the tool for grinding lens 1. During grinding the tool can be supported between blocks attached to any rigid work surface. Abrasive slurry is applied to the glass. Lens 1 is placed on top of the slurry and ground by conventional center-over-center strokes. The excursion of the strokes should be about a third of the diameter of the glass blanks.

"The radius of curvature can be monitored by the same technique used for checking the primary mirror. If the radius becomes shorter than desired, simply reverse the position of the blanks: turn the pair upside down and grind lens 2 on lens 1 to increase the radius by the desired amount. This procedure may be necessary during the later stages of fine grinding, although experimentation will demonstrate that some control of the radius can be exercised by altering the length of the grinding stroke. Strokes longer than about a third of the diameter of the blanks tend to decrease the radius of curvature; those shorter than a third of the diameter increase the radius. The lens is designed to be achromatic. The image will be free of spurious color if the radius of these curves does not depart from the specified dimension by more than .02 inch.

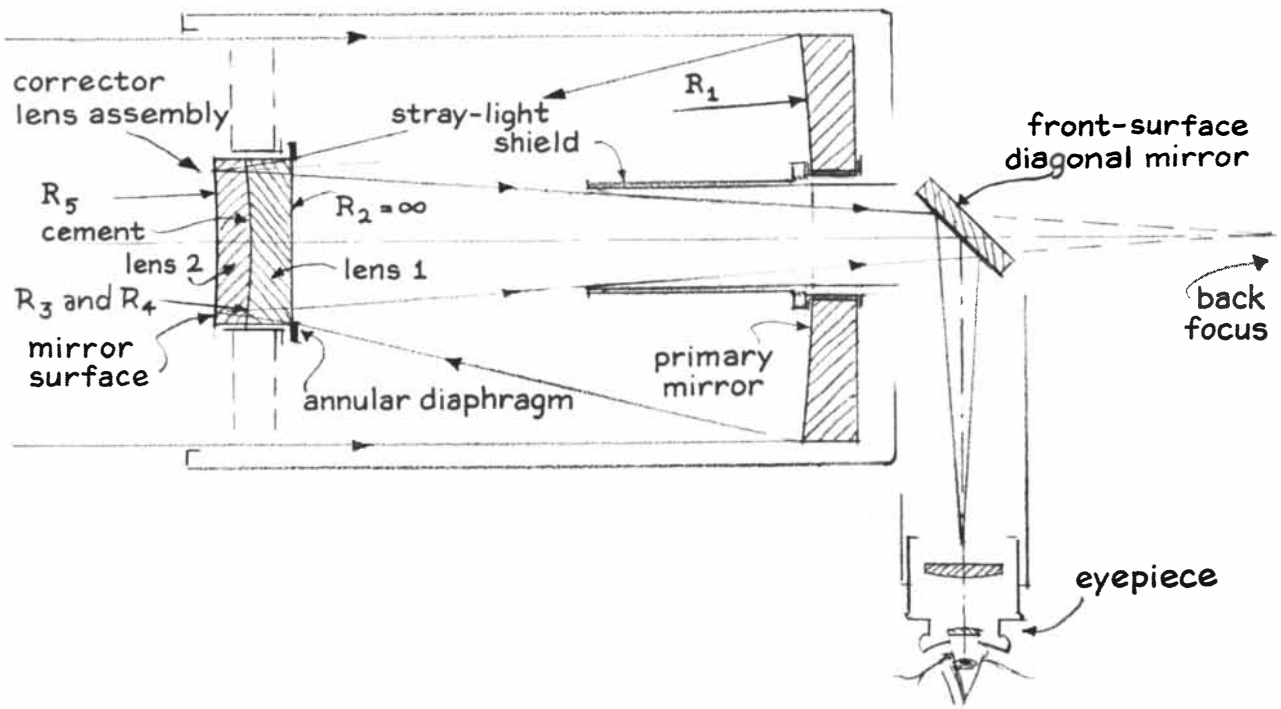
"I next ground and polished R_5, the external surface of lens 2. This surface is ultimately aluminized to function as the secondary mirror of the telescope. An extra disk of glass, which is eventually discarded, is used as the grinding tool. The radius should be kept to within .2 inch of the specified dimension.

"Care must be taken when grinding R_5 to keep the two surfaces of the glass concentric. Rotate the glass as grinding proceeds in order to prevent the lens from becoming wedge-shaped. Monitor the thickness frequently by measuring the edge with a dial micrometer. The final thickness of the polished glass should be within .025 inch of the specified dimension.

"Fortunately the thickness of the separate elements of the corrector-lens assembly is less critical than their sum. An error of thickness in lens 2 can be corrected by an opposite change in the thickness of lens 1. Incidentally, the thickness of lens 1 is somewhat easier to control than that of lens 2 because the external surface of lens 1 is flat. It is almost impossible to avoid grinding a small wedge angle into lens 2.

"Wedge error can be described by saying that a line connecting the centers of curvature of the two surfaces will not pass exactly through the center of the assembled lens. The object is to keep the error as small as possible. The effect of small errors can be minimized when the system is assembled by masking off the periphery of the combination so that rays of light pass through only the symmetrically thick portion of the glass.

"The usefulness of a dial micrometer and a turntable for checking the radius of curvature and the tendency of the glass to become wedge-shaped can scarcely be overstated. I used a small lathe as the turntable. A series of eight marks spaced at equal intervals was



Arrangement of lenses in Robert J. Magee's optical system

Component	Diameter (inches)	Radius of curvature (inches)	Index of refraction	Dispersion	Thickness (inches)
Primary	4.25	$R_1 - 19.25$	(mirror)	—	.625
Space between primary mirror and lens 1	—	—	—	—	5.8
Lens 1	1.75	$R_2 - \infty$ $R_3 - 6.88$	1.517	64.5	.489
Cement	—	—	(cement)	—	0
Lens 2	1.75	$R_4 - 6.88$ $R_5 - 12.75$	1.649 (mirror)	33.8	.297
Back focal length (between vertex of R_1 and focal plane)		5.032 inches			
Effective focal length of the system		33.47 inches			

Optical dimensions of the system

made around the edge of the lens with waterproof ink. The marks serve as references for clamping the lens in the lathe consistently at the same orientation with respect to the jaws of the chuck and also for identifying regions of the glass that require additional grinding.

"To check the tendency of the glass to become wedge-shaped seat the lens firmly in the jaws of the chuck and clamp it lightly. By manipulating the transverse and cross feeds of the lathe place the contact point of the micrometer against the glass near the edge of the lens. Rotate the chuck slowly. The pointer of the micrometer will remain stationary if the thickness of the lens is uniform.

"The diameter of the lens and the depth of the sagitta, or concave surface, can also be measured with the micrometer and the turntable. When the depth of the concave surface is known to within .0005 inch, the radius of curvature can be computed to within .25 inch. Dial micrometers can be read easily to within .0002 inch.

"The accompanying diagram [*top of opposite page*] gives the geometry of the lenses. The sagitta is equal to the radius minus the square root of the difference between the square of the radius and the square of half of the diameter of the lens, or $S = R - [R^2 - (L/2)^2]^{1/2}$, in which S is the sagitta, R is the radius and L is the diameter of the lens. For example, the sagitta of a lens 1.75 inches in diameter that has been ground to a radius of 12.75 inches is equal to $12.75 - (162.5625 - .7656)^{1/2} = .03005$ inch.

"This formula enables the worker to anticipate the depth of the curve that will be needed to achieve a desired radius. Conversely, I have found it helpful to measure the sagitta periodically during the grinding operation and compute the diminishing radius as the work proceeds. The formula is $R = L^2/8S + S/2$. For example, a lens 1.75 inches in diameter when ground to a sagittal depth of .03005 inch has a radius of $3.0625/.2404 + .03005/2 = 12.75$ inches. Make the measurements with care when working with lenses of these proportions.

"Generating the flat surface of lens 1 will doubtless require the most patience. The project will be simplified if the worker has access to a standard optical flat of the same diameter as the lens or larger. The polished surface of the lens is tested for flatness against the standard by optical interference. To make the test place the standard on a solid support with the flat side up. Rest the lens, flat side down, on top of the standard. Separate the pair at one side by inserting a piece of tissue paper between the glasses near the edge.

"Flood the pair from the top with monochromatic light, such as the yellow rays emitted by the flame of a gas burner that plays on a wick moistened with brine. Examine the reflected pattern of light. It will consist of a grid of light and dark bands known as interference fringes. If both optical surfaces are flat, the fringes will form a grid of straight, parallel bands that are alternately light and dark. Curved fringes indicate departure from flatness.

"The depth of the curvature is determined by placing a straightedge across the center of the lens in a position that joins the ends of a complete fringe by a straight line, as the bowstring connects the ends of a bow. Multiply by 12 the number of partial fringes that are enclosed by the complete fringe and the straight line to determine in millionths of an inch the approximate deviation of the surface from flatness. The accompanying diagram [*bottom of opposite page*] depicts the interference pattern generated by a surface that departs from flatness by approximately 36 millionths of an inch, or three fringes.

"The external surface of lens 1 must be ground and polished to within less than half of a fringe of flatness. A simple method of generating a flat surface requires three glass disks of equal diameter, one of which can be the lens. The other two disks should be at least a quarter of an inch thick. The procedure is based on the principle that if three surfaces consistently make full contact

when placed together in every possible combination, all must be flat.

"Number the edge of each disk with waterproof ink. Begin by grinding disk 1 on 2, then 2 on 3, then 3 on 1. Next, invert the sequence by grinding 2 on 1, 3 on 2, and 1 on 3. Return to the first sequence and thereafter proceed alternately. Use conventional center-over-center strokes about .3 inch long. Grind with a slurry of No. 600 grit in water. Limit the grinding to 25 strokes per pair of surfaces and continue until all surfaces have been fully ground. Finish with 10 strokes per pair of surfaces.

"Prepare a polishing lap by coating one of the glass disks with hot pitch. When the pitch cools, divide the lap into facets about .3 inch square by cutting grooves in the pitch. Coat the ground surface of the lens with a slurry of rouge in water, place the coated surface on the lap and apply about a pound of pressure for 30 minutes, or until the facets of pitch flow into full contact with the glass. Polish the lens for about 10 minutes.

"Test the incompletely polished surface against the standard flat without inserting tissue paper at the edge. If the grinding has been carefully done, no more than one or two concentric fringes will be observed. The fringes indicate that the surface of the lens is either uniformly concave or uniformly convex.

"Exert downward pressure on the edge of the lens. If the surface is convex, the fringes will move toward the point where pressure is applied. If the fringes do not move, exert pressure on the center of the lens. If the lens is concave, the radius of the fringes will increase. To correct the curvature, cut pitch from the edges of the grooves to reduce the area of the polishing facets uniformly toward the center or toward the edge of the lap as needed to flatten the surface.

"Continue polishing with the modified lap. Test the surface frequently as polishing proceeds and alter the area of the facets as required. If the correction is carried too far or if irregular zones develop, make a new lap. Continue until the surface is fully polished and flat. The operation is not as difficult as it may seem. It requires more patience than skill. If a standard optical flat is not available, one can be made in a matter of hours by the procedure described in *Amateur Telescope Making—Book One,* edited by Albert G. Ingalls (Scientific American, Inc., 15th printing, 1962).

"After all five surfaces have been ground, polished and measured, finish cutting the hole through the center of the primary mirror. Before cutting the hole, cover (and thus protect) the polished surface with a sheet of paper coated with pressure-sensitive adhesive. Paper so coated is available from dealers in art supplies. The primary mirror and the concave surface of lens 2 can now be aluminized. If the secondary mirror is aluminized, it is worthwhile to have the flat surface of the corrector coated for low reflection, since this surface is used twice. Usually a company that does aluminizing will also do coating.

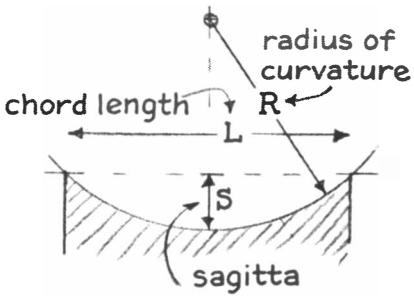
"The task of making the optical elements is finished by cementing the mating surfaces of the corrector lens. This job is simple in principle but difficult in practice. The lenses must be thoroughly cleaned, preferably in a solution of nitric acid or trisodium phosphate. Airborne dust must be excluded from the work area.

"The lenses are heated to 150 degrees Fahrenheit in a water bath, dried and cemented at this temperature. Lens 1 is supported, flat side down, on a tabletop covered with a sheet of lens tissue. A few drops of warm Canada balsam are poured in the center of the concave surface. The mating surface of lens 2 is gently lowered into contact with the cement. A downward pressure of about two pounds is exerted on the assembly for several minutes to squeeze out the excess cement. The excess can be cleaned from the edge of the lenses by a cloth moistened with xylene.

"A short, snugly fitting tube is slipped over the combination to keep the lenses centered. A pad is placed over the aluminized surface of lens 2 and a two-pound weight is placed on the pad. The cement will set in about two days.

"The job may not go easily. Bubbles that are difficult to remove tend to become trapped between the lenses. It may be necessary to alter the viscosity of the cement. For these reasons the inexperienced worker is urged to review the portion of the article on lens making by J. R. Haviland in *Amateur Telescope Making Advanced—Book Two,* edited by Ingalls, that describes some of the tricks of using Canada balsam cement.

"Short telescopes that employ two reflecting mirrors require a carefully designed system of baffles to improve image contrast, or at least to preserve it, particularly if the instrument is to be used in daylight. The reason is not hard to find. Assume that a bundle of rays enters the telescope at an angle such that it just grazes the corrector lens, passes through the hole in the primary mirror and illuminates the focal plane. The rays

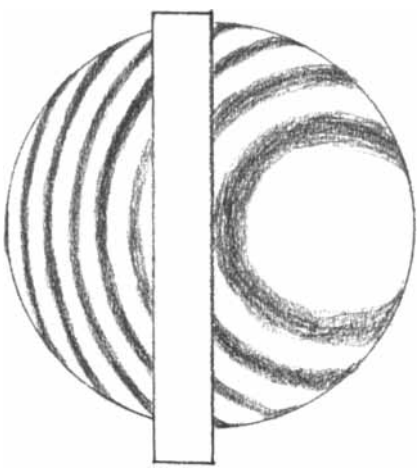


Geometry of spherical lens

make no contribution to the image because they have not fallen on the face of the primary mirror and hence are not focused. They simply flood the image as a veiling glare.

"Such a glare cannot develop in a telescope of the Newtonian type that includes a solid tube because the diagonal mirror that diverts light into the eyepiece faces away from the incoming rays. Only the rays that are reflected by the primary mirror fall on the diagonal mirror. Instruments that have a pair of reflecting mirrors such as those in my design can be effectively baffled in several ways. The combination of a tubular shield that extends forward from the center of the primary mirror and an annular diaphragm surrounding the corrector lens is effective if the parts are properly proportioned. The object is to make the tubular light shield sufficiently long and the annular diaphragm sufficiently wide to prevent rays from sources beyond the field of view from reaching the focal plane directly.

"The baffles are installed at some cost in terms of the brightness of the image. The shields block out the central portion of the mirror more or less depending on



Fringes indicating spherical surface

How some oil companies view it.

How we view it.

Some things have been around for a long time, just waiting to be seen. When somebody looks hard enough to notice them, they'll be considered important new discoveries.

When we look at crude oil, we always try to look for something new . . . for the kind of things other oil companies don't always think to look for. That's how we can keep coming up with newer and better oils for your industry.

At American Oil, it's more than what we look at . . . it's how we look at things that makes us important to you.

AMERICAN

American Oil Company

TV STATION MASTER CONTROL ROOM

ERICH HARTMANN/MAGNUM PHOTOS

THE SEPTEMBER ISSUE OF **SCIENTIFIC AMERICAN** WILL BE DEVOTED TO

COMMUNICATION

their proportions. On the other hand, the center of the mirror does not work anyway because it is perforated and blocked by the corrector-lens assembly. The short tubular shield that extends forward from the primary mirror is equivalent as a baffle to extending the main tube of the instrument about two feet.

"Incidentally, the shielding tube can be extended behind the mirror and clamped mechanically by the end plate of the main tube. It then serves as a peg for supporting the primary mirror. The diagonal mirror can also be supported in the shielding tube. A window cut in the side of the tube admits rays to the eyepiece. I omit the details of the mechanical construction because they will vary with the builder's taste and the contents of his scrap box.

"The 4¼-inch Pyrex blank for the primary mirror and the lens cement can be bought from the Edmund Scientific Co. (Edscorp Building, Barrington, N.J. 08007). A number of companies aluminize mirrors, including the Research Service Company (1149 Massachusetts Avenue, Arlington, Mass. 02174). Optical glass of the quality required for the corrector lenses is available from A. Jaegers (691A Merrick Road, Lynbrook, N.Y. 11563). Enough glass for three sets of elements can be cut from one piece each of the glasses listed respectively as Catalogue No. 1590 (for lens 1) and 1591 (for lens 2) in the Raw Optical Glass section of this company's catalogue. Incidentally, when you reduce the thickness of these glasses, care should be taken to leave ample margin for correcting errors of curvature and wedge angle.

"When you assemble the telescope, take pains to align the optical axis of the primary mirror with the optical axis of the corrector lens. Misalignment is known as decentering. A decentering error of .005 inch can be detected but is tolerable.

"I designed the optical system by the technique of ray tracing. This mathematical procedure can be rather tedious when it is done with pencil and paper. Fortunately desk computers are available. With such a computer it is possible to adapt to electronic computation the ray-trace equations as given in textbooks on optics. It is important, however, to find a machine that has keyboard trigonometric functions. A desk computer is not difficult to operate, although I urge the programmer to arrange his calculations in a systematic manner and to incorporate a means of verifying his answers."

BOOKS

The eye-brain, hovercraft and hydrofoils, 600-m.p.h. automobiles and other matters

by Philip Morrison

OPTICS, PAINTING & PHOTOGRAPHY, by M. H. Pirenne. Cambridge University Press ($17.50). FOUNDATIONS OF CYCLOPEAN PERCEPTION, by Bela Julesz. The University of Chicago Press ($20). The human cortex is heavily committed to the analysis of visual images, an analysis that we know very well makes intelligent viewing an act far less passive than is suggested by the familiar metaphor of the camera. Just as Christopher Isherwood's novelistic sensibilities transfigured Berlin to make "I am a camera" a rhetorical figure, so does our eye-brain, largely at the unconscious level, transform its optical input to give us the visual world we "see." Indeed, as M. H. Pirenne says in the first of these books, "the eye is the only optical instrument which forms an image which has never been intended to be seen." The retinal image is inverted, hemispherical, constantly moving. The two books treat differing aspects of that circumstance, the first a topic as cave-old as painting itself, the second a novel means of studying the visual act.

Pirenne, the son of a painter and himself an amateur painter, is a distinguished Oxford physiologist and a major contributor to our knowledge of vision. This compact, clear, rather elegant work is addressed to serious readers who are artists and photographers, who want to understand how we make pictures. The book is not mathematical, although it does not skimp on geometric argument. Its ample illustrative material and lucid historical accounts enrich the entire presentation, making the book readable even though it is by no means merely introductory.

The center of the book lies in a formidable church in Rome. In that severe structure of darkened stone there is a visual wonder. It is a great fresco depicting St. Ignatius being received into heaven, painted on the hemicylindrical vault over the nave by Fra Andrea Pozzo, S.J., at the end of the 17th century. (The fresco is given one star in the green *guide Michelin*, which dryly remarks "Trés baroques.") The highest point of the vault is 30 meters above the floor. The viewer sees the blue sky with fleecy clouds, peopled with angelic and human personages, framed in a piece of imaginary vaulting architecture that appears to carry the real walls of the church far aloft. Pozzo projected his entire conception onto the ceiling with a net of squares stretched across the nave and threads carrying the viewpoint from a central yellow marble disk on the floor up to the vault surface. "The result of all this work is striking.... The spectator is unable to see the painted surface, *qua* surface.... The arches... are seen to stand upright into space. They are seen in three dimensions, with a strength of illusion similar to that given by the stereoscope.... The whole ceiling looks like a vast complicated piece of coloured sculpture... a species of art quite distinct from ordinary painting."

The illusion is imperfect, of course; the painted sky cannot be bright enough. For depth, however, it is "well nigh complete." The flux of light at the viewing disk is almost what it would be if the painting were replaced by the colored sculpture of the scene. Photographs show no such illusion, and a viewer who walks away from the yellow disk sees the structure deformed into one that seems in danger of collapse. The point is plain: the single eye, viewing in central projection, sees a true linear perspective view, just as it would see the real structure. (Binocular cues are ineffective because the scale is too large.) Normal pictures and photographs are not so faithful to the optical laws of perspective. We see depth in them only to the degree that our brain interprets the picture surface; they suggest depth, but they do not literally present it.

There is an arcade in the Palazzo Spada, built by Borromini about 1650, that inverts this argument, as the distorted Ames rooms familiar in this country for a couple of decades do. The designer arranged his columns and all other solid portions to cover accurately, for a spectator at the entrance, an imaginary larger arcade of constant width and height. The real arcade decreases in height and width as you move along it, so that it looks much longer from the entrance than it actually is. A man standing at the end of the arcade looks like a giant, since one way we learn sizes is to accept perspective comparisons whenever they occur in familiar perceptual context. Let the picture surface once appear and entirely new and largely unknown rules govern.

The story is made clear by examples of the compromises and adaptations artists have made to weave their fabric of space out of the geometric constraints of perspective projection, with the freeing power of the viewing mind. A mirror presents the eye with not an illusion but approximate reality. (The mirror image is a wide-band one; a hologram achieves the same result in a narrow-band image.) The test is that the entire scene can appear even if you view only a small part of the mirror, even if you vary direction and position. No picture can do this; our satisfaction with pictures arises not from optics but from the stored experiences and cunning processes of the mind. Yet the perspective rules are no mere Renaissance conventions; they are faithful, though confined, maps of the real geometry of light.

Dr. Julesz, who studies perceptual processes at the Bell Telephone Laboratories, has found a new style of *trompe-l'œil* undreamed of by painters and part of our digital times. His "pictures" are familiar to readers of this magazine, where they have been presented in the past. Julesz prepares two frames, in neither one of which can the viewer see anything but a textured pattern of random dots. Fuse the frames by any means of stereoscopic vision, one to each eye, and with wonderful slowness the "cyclopean eye" within—the eye of the mind—will compute from them the subtle correlations that display a form in vivid depth. There is a processing delay of many seconds, a pregnant sign of the interior mind. "Cyclopean" vision lies

First price is publisher's list. **Boldface** shows member's price.

736. The Quest for Theseus
Edited by Anne G. Ward. The historical reality behind the legends of the mythic slayer of the Minotaur and founder of Athens. $13.50/**$8.95**

885. The White Generals: The White Movement and the Russian Civil War
By Richard Luckett. $10.00/**$7.50**

859. Let History Judge: The Origins and Consequences of Stalinism
By Roy A. Medvedev. $12.50/**$8.50**

858. Napoleon Bonaparte: An Intimate Biography
By Vincent Cronin. $12.50/**$8.50**

857. The Rise and Fall of the Habsburg Monarchy
By Victor-L. Tapié. $15.00/**$8.95**

677. Witchcraft at Salem
By Chadwick Hansen. $6.95/**$5.75**

754. Augustus to Constantine: The Thrust of the Christian Movement into the Roman World
By Robert M. Grant. $10.00/**$7.50**

726. A History of the African People
By Robert W. July. $15.00/**$8.95**

796. Justinian and Theodora
By Robert Browning. The peasant's son and lady of ill-repute who ruled the Byzantine Roman empire. $15.00/**$9.95**

105. Huey Long
By T. Harry Williams. National Book Award & Pulitzer Prize. $12.50/**$8.50**

877. Apache Chronicle: The Story of a People
By John Upton Terrell. $12.50/**$8.50**

876. Rebirth of Feminism
By Judith Hole & Ellen Levine. $10.00/**$7.50**

775. The Rising Sun: The Decline and Fall of the Japanese Empire, 1935-45
By John Toland. $12.95/**$8.95**

832. China and Russia: The "Great Game"
By O. Edmund Clubb. Three centuries of "massive contest," always short of war. $12.95/**$8.75**

103. From Slavery to Freedom: A History of Negro Americans
By John Hope Franklin. $13.95/**$8.40**

746. Empire of the Steppes: A History of Central Asia
By René Grousset. $17.50/**$9.95**

891. The Double-Cross System in the War of 1939 to 1945
By J. C. Masterman. How the British actually controlled and used German spies. $6.95/**$5.95**

460. Hitler: A Study in Tyranny
By Alan Bullock. $10.00/**$6.95**

759. Gold, Glory, and the Gospel: Lives and Times of the Renaissance Explorers
By Louis B. Wright. $10.00/**$7.50**

476. The Life of Lenin
By Louis Fischer. Winner of the National Book Award. $12.50/**$8.50**

606. Henry VIII
By John J. Scarisbrick. $10.95/**$7.65**

671. Medieval History: The Life and Death of a Civilization (2nd Edition)
By Norman F. Cantor. $10.95/**$8.50**

687. Anti-Intellectualism in American Life
By Richard Hofstadter. Pulitzer Prize winner. $7.95/**$5.95**

657. The Creation of the American Republic, 1776-1787
By Gordon S. Wood. Bancroft Award winner. $15.00/**$9.95**

The History Book Club
Stamford, Connecticut 06904

Please enroll me as a trial member and send me, for only 99 cents, the three books whose numbers I have filled in:

[ ] [ ] [ ]

Also send me, at the special low member's price:

[ ]

I enclose no money now. Send bill with books.

Within two weeks, I may return the books and owe nothing. Or keep them and agree to take only four more selections in the coming year from among more than 150 offered each month in the Club Review. I may choose any of these instead of the Editors' Choice, or no book at all, by returning the reply card accompanying the Review. SA 08S

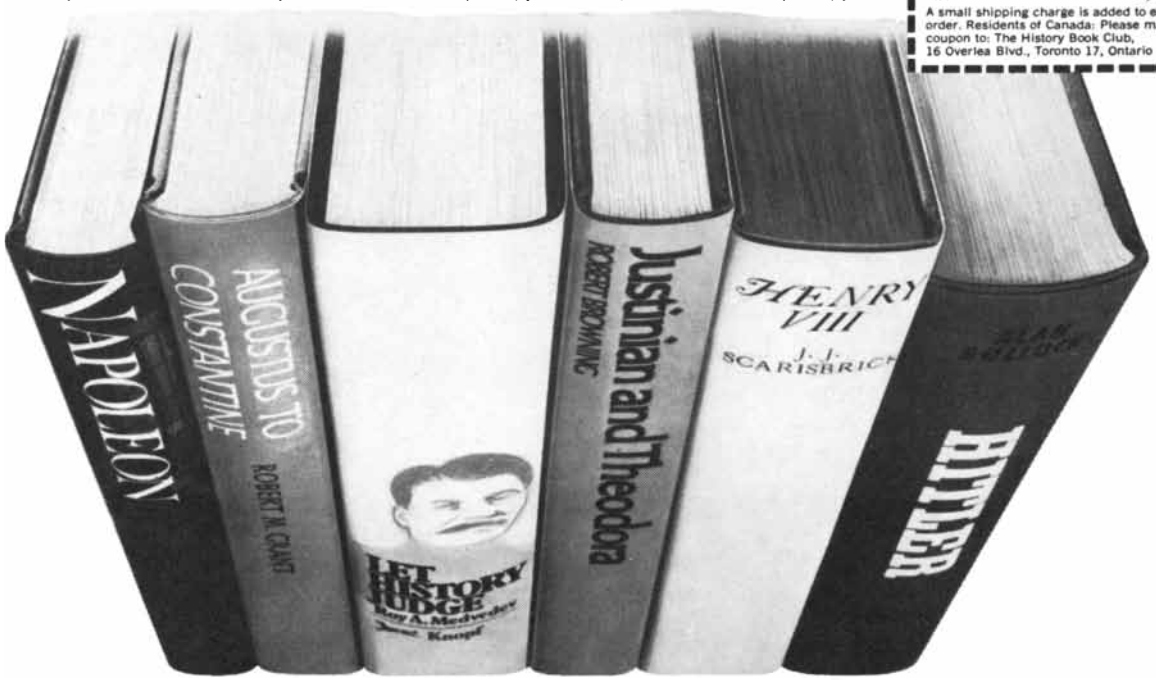
Print Name

Address

City

State Zip

A small shipping charge is added to each order. Residents of Canada: Please mail coupon to: The History Book Club, 16 Overlea Blvd., Toronto 17, Ontario



Special on tyrants. Any three for 99 cents.

Hitler. Stalin. Napoleon. Augustus. Lustful Henry. Justinian and his Theodora.

They did more to change the course of history than many of the good guys.

That's why we'd like to send you any three of these gentlemen as your introduction to The History Book Club.

Or, if you prefer, you may choose the company of heroes, martyrs, politicians, kings. Or become embroiled in wars or revolutions.

In fact, you can have any three books listed above. All for 99 cents, when you take a fourth at the low member's price.

If you're not delighted, you may of course return them. But if you do keep them, you simply take four more new books during the coming year, from the 150 offered each month. You get these at savings of $2, $3, or $4 off bookstore prices. And you also earn other savings through bonus books you can choose from the Club's entire list.

Well-known statesmen, scholars and community leaders belong to The History Book Club. But by accepting our tyrants offer, you can also enjoy the company of villains.

The History Book Club

Stamford, Connecticut 06904

within the brain, but the term is meant to imply not only centrality of location but also the system's capacity to be stimulated by percepts.

Given this remarkable result, the effect is employed to examine many other problems in visual perception. Most simple optical illusions—a line with arrowheads at the ends looks shorter than the same line with the arrowheads reversed—are generated by the Cyclops within. Ambiguous figures can be made to order, so that you can view more or less at will either the front view of an object or the back view. The cyclopean eye will even put sparse dots together into a floating surface. It can get depth from patterns of line segments that differ in direction, and patterns where black and white alternate. One can measure how long it takes the visual system to focus on, fuse and compare the two patterns. There are aftereffects and fatigue effects. There are some remarkable dynamic properties, learned from computer-constructed motion pictures. Each eye alone sees randomly stirring dots. With two eyes a square stands out in depth, perhaps wiggling. The study can be extended to spatial frequencies—bar patterns of different spacings—with results that do not make it quite clear whether stereopsis takes place before the judgments of spatial frequency or afterward.

About 100 stereograms are presented (both as image-fusion pairs and in the more easily accessible red-green anaglyph style), so that the reader can examine the main issues for himself. Only the motion pictures are lacking; we are promised them in 16-millimeter form by the publisher.

Julesz' book is quite different from Pirenne's. There is not much more mathematics present, but the work is intellectually embedded in the neurophysiological and computer-based perception studies of the day. Eigenvalues and transformations (Fourier and otherwise), noise measures, simplified algorithmic models for one effect or another and a somewhat elaborate analogue model (with interacting magnetic needles mounted on a spring mattress) all turn up. The general reader will be led to skip chapters; the reader with more specialized commitment to modern psychology cannot really afford to do so. Everyone will find that time spent looking at the diagrams and musing on the main arguments is well spent.

The most surprising topic—it plainly surprises and intrigues the author greatly—is the certain report that at least one subject possesses a vivid eidetic memory, which she can use to catch the cyclopean eye. This young woman, an artist, can view a random-dot pattern with one eye one day and its matching stereoscopic pattern the next and then fuse them in the cyclopean eye from memory. She can name and describe the correct pattern, and she retains her abilities through a long set of controls. If she is shown as many as four patterns, each in a distinct color, she can store all four and reclaim the stereoscopic effect one by one according to color! All of this work is done with 100-by-100 arrays of dots, although it is not clear how much information from the dot pattern needs to be retained in order to allow correct identification.

Says Julesz: "These experiments appear incredible." The eidetic memory clearly occurs "in front of" the cyclopean "retina" for stereopsis. No one believes that the anatomical retina could possibly be given information on any large scale from the mind; there are few, if any, outgoing pathways. Therefore, Julesz conjectures, the eidetic image is formed later than the image on the retina but earlier than the stereoptic image, perhaps much earlier. We are perceiving layers of the visual mind. What is needed is more such remarkable individuals; the book includes a plea for a search among "young people, artists, and other special groups with high yield." It also includes a test with only some 60 dots for screening eidetikers. Those who pass this simple test are eligible for the eidetic finals!

The scent of progress is strong here. From both of these books arises the hope that we are on the right track in our search for how it is the human mind can perform its still unnumbered functions. The goal nonetheless lies far head.

JANE'S SURFACE SKIMMERS: HOVERCRAFT AND HYDROFOILS 1971–72, edited by Roy McLeavy. McGraw-Hill Book Company ($30). LAND SPEED RECORD, by Cyril Posthumus. Crown Publishers, Inc. ($9.95). Sojourners along the lovely Yugoslavian beaches can almost every day watch a sea monster whizzing across the blue horizon. What else could it be? It is too swift, too light, for any ship; it is poised in the air as though it were walking on the waves. All the same it is a working diesel-powered vessel, one of the trim 100-passenger Kometa hydrofoils built in shipyards along the Volga. Its entire welded-aluminum hull rides high above the smooth sea on strong struts; it "flies" on submerged foils at some 26 miles per hour. Each summer this model and its kin haul millions of passengers across the inland waters of the U.S.S.R.: down the arctic river Irtysh and across the desert-locked Caspian. The Yugoslavs run a flotilla of them on their Adriatic coast; others ply the Danube and even the Thames. Now, in the latest air-conditioned version, they traverse American harbors in the Virgin Islands. There is a definite commercial-hydrofoil gap; the Russians far outbuild the runner-up, an experienced Italian firm on the Strait of Messina. The largest hydrofoil is American, but it is a naval prototype. It is built by Lockheed on Puget Sound and is operated there by the Navy as a test of seagoing warships to come. It can drive its 320 tons at 60 m.p.h., and with a projected doubling of its twin gas-turbine power plant it should eventually reach 100 m.p.h.

If such inverted wingborne travel should pall, there are hovercraft aplenty. These vessels ride on a cushion of pressurized air held in a big chamber edged with a flexible "skirt," often an inflated torus of rubberized fabric. They are usually driven by airscrews. The biggest hovercraft are the Channel automobile ferries of the British Hovercraft Corporation, which ply that choppy run hourly in the peak season, turning up 50 m.p.h. in waves 10 feet high with a load of some 30 cars and a couple of hundred passengers. Abovedeck these craft of the Mountbatten Class (four or five are now in service) sprout four big four-blade aircraft propellers. There are also racing, heavy utility, rescue, police, even kit-built hovercraft, their builders all over the world. Here they are described in that knowing and exhaustive style dear to *Jane's*, with plenty of photographs and plans and not a few drawings out of the dreamy future. One working example will have to suffice: the English Hovertrailers, a range of "non-self-propelled" air-cushion vehicles that will pull heavy loads over "marsh, bog, mudflats, ploughland and construction wayleaves." On a rectangular steel platform there is a welded subframe carrying a segmented skirt; low-pressure air supplied by a diesel-powered fan bears the load lightly in models of up to 50 tons. The drawbar pull is low—on level ground a twentieth or less of the load—and a 10-ton trailer can "easily surmount a 3 feet bank at 45° under full load."

Propeller-driven cars that hover inches above their special tracks need only a light support structure because of the absence of any concentrated stress. Many experiments in intercity and urban transport are under way based on this concept, air "rails" being even handier than the old monorail. American proto-

types were shown recently at the Dulles International Airport exposition. The most developed of these systems is the Bertin Aerotrain for the Orléans-Paris run, 70 miles at 155 m.p.h. "in airliner comfort" on a handsome prefabricated concrete-beam track 16 feet above the ground. Much practical testing has been completed over the first 11 miles of track; commercial service lies ahead.

Jane's editors, informed if enthusiastic, see early "lift-off" for the entire surface-skimming industry. Central to it will, they think, be naval craft; the way will be led by the U.S. Navy, worried about budgets yet anxious for speed and flexibility. Mile-a-minute, water-jet-driven hydrofoil patrol boats, following a well-tested U.S. prototype built by Boeing, will be used by all the NATO countries. After that are forecast "small, high speed ocean-going weapons platforms," as big as a destroyer and bigger and moving at 100 knots. Even stranger vehicles—looking like big half-jet, half-turboprop flying boats but designed to stay close to the sea surface to get cheap extra lift from their moving aerodynamic image below, called "wing-in-ground-effect" machines—are projected, somewhat mistily. The Russians are said to be trying out such a craft as a troop carrier for 800 men at 200 m.p.h.! Skirts and static air cushions are slow and skittish by comparison, although they may be right for tundra and ice pack. The crude displacement of water will be left to submarines and tankers.

Speed on land—for sport and fame—is the strangely pure motif of *Land Speed Record,* whose author is a well-known journalist in the world of automobiles. It traces in narrative detail, quite nontechnically but with knowledge and care, the entire story of the quest for the world record in automotive speed, year by year since the first *course de vitesse* near Paris in late 1898. The first "flying kilometer" record speed of 39.245 m.p.h. was set by the Comte de Chasseloup-Laubat in an electric car, chain-driven by a 36-horsepower series-wound high-torque electric motor working on a heavy load of nonrechargeable batteries. It is worth noting that this speedster, "trailing a few blue sparks and a slight tang of ozone," fell just short of the bicycle record for the flying kilometer, which then stood at 40 m.p.h. Things have changed. Tracks gave way to beaches, and beaches were sought the world over for wide, long, safe straightaways. Dry lakes and desert hardpan were tried; no one knows a more practical site in all the world than the dozens of hard, smooth miles of the Bonneville flats in Utah, where in 1909 the railroad builders brought a waterline 35 miles from the mountains to make possible the town of Wendover, at the edge of the glaring white salt of the evaporated glacial lake bed. Every record since 1935 has been set there, save one on dry Lake Eyre in central Australia.

Today records are dual. The wheel-driven record, set in 1965, is 409 m.p.h. The golden car that did the job still looked like an automobile; the jet and rocket vehicles that hold the absolute speed records in their special class tend to resemble fighter aircraft whose builders forgot the wings. *Goldenrod* was even powered by automobile engines: four boosted near-stock big Chrysler V-8's. The enclosed cockpit was behind the rear wheels; the entire 32-foot car—nearly all hood—was barely knee-high, except for a few humps for the wheels and air intakes and a tall fin that enclosed the braking parachute. *Goldenrod*'s record stands at 409.277 m.p.h., even though Bob Summers, who drove the car he and his experienced hot-rodding brother Bill designed and built, never got past third gear. Tires driven by 2,400 horsepower, however, are now outmatched. *Blue Flame,* rocket-propelled by liquid natural gas and hydrogen peroxide reacting in a carefully designed engine, its power equivalent to 35,000 horsepower during burn, riding on paper-smooth tires inflated to 350 pounds per square inch, holds the absolute record (unofficially peaking ahead of even J. P. Stapp's railborne rocket sled) at 630.388 m.p.h. "Its attendants resemble scientists and doctors rather than the oil-smeared motor mechanics of earlier days." Yet some experiences are invariant: the joyful photograph of Gary Gabelich, the driver, with the three heads of the enterprise right after the run is timeless. The designers came out of the drag-racing world, with a strong assist from advanced professional engineering circles, and "over seventy lucky students were able to join in on the design side of the work" at the Illinois Institute of Technology. The record was set in the fall of 1970; *Blue Flame* has plenty of thrust left to unthrottle, and the push for Mach 1 and beyond, all the way up to 900 m.p.h., seems clearly ahead.

The classic, colorful airbrush drawings on a white background are here for about 60 beauties (and a couple of monsters), aspirant and winner alike, ever since 1899. The first challenger, *La Jamais Contente,* an electric car with two direct-drive motors that took the record away from the Comte de Chasseloup-Laubat in 1899 at a swift 65.79 m.p.h., is still the flower of them all: a silver-blue double-pointed aluminum-alloy spindle. The artist who made the drawings is Michael Roffe.

OXFORD ECONOMIC ATLAS OF THE WORLD, FOURTH EDITION, prepared by the Cartographic Department of the Clarendon Press. Oxford University Press ($25). Forty-six full-page maps of the world, some 15 inches by eight, many in handsome color lithography, make up the bulk of this much enriched version of a well-known book. With them are many smaller world maps and a virtual mosaic of detailed inset maps. On each map is plotted some economic distribution of interest. These include data on the physical environment, crops, minerals, energy use and manufacturing and trade in many specific commodities (not only iron ore and bananas but also aircraft and computers), plus social information such as demographic statistics, disease incidence, estimates of education and medical resources. There are maps of transport nets on land, sea and air and an economic gazetteer with a thick set of tables gathering the numbers country by country.

Economic atlases have intrinsic troubles. Ordinary geography, even at borders, is much less transitory than economic data. For example, we see here the cleverly plotted flow of oceanic shipping. The atlas-makers chose from Lloyd's list a sample of ships, located each of them on one particular day early in 1967 and drew a band of color along each sea route adjusted to a width proportional to the tons of ship displacement moving per mile of route. Then Suez closed down a few months later; in fact, another map, presenting crude-oil flow, shows the oil now surging around the Cape of Good Hope. Four thick sea arteries carry most of the black nutrient fluid; the main one runs around Africa from the Persian Gulf to the English Channel, the second from that same hot oily gulf (now via Borneo and Singapore) to southern Japan, the third from Libyan and Algerian ports to southern France and the Genoese coast and the fourth from the Texas Gulf ports to the harbors of our Northeast. A parade of laden tankers is strung along these main routes, one every 50 or 75 miles, day and night the year round.

Symbols for quantitative data are hard to make truly graphic. The problem is solved here about as well as it is anywhere, although some of the maps are not easily legible. Color and form are

artfully used to help distinguish the wide ranges most economic quantities require in this unequal world. The data are given mainly for the period between 1954 and 1964, allowing for some estimate of growth (although it is a little less timely than one might hope). Statistics for a number of countries, particularly China, are rather incomplete, a fact not made as clear as it might be on each map. These are cavils; no work of the kind in English is better, and surely none is more attractive in maps, letterpress and binding. The price is not low.

Here are a few gleanings from the volume: the density of cattle is nowhere greater than it is on the Ganges Plain, where the dots run together over a 1,000-mile stretch. That is no surprise, but the second most noticeable concentration of cattle is on the plains of western Ethiopia! The universities of Finland, the U.S.S.R. and the West Indian islands have the largest proportion of women students. (To be sure, Burma, the Philippines, Cyprus, Israel and New Guinea do as well, but the symbols concentrate notably in the narrow Caribbean.) Leprosy is still a scourge in West Africa. The brewing of beer and the manufacture of cement are very widespread industries; their local presence is an early sign that a country is beginning to industrialize. And of all the names in the book the site of the Malagasy Republic thorium and rare-earth deposits rolls most sonorously off the tongue: Ambindandrakemba.

McBURNEY'S POINT: THE STORY OF APPENDICITIS, by Stewart M. Brooks. A. S. Barnes and Company ($6.50). Draw a line from the navel to the frontal protuberance of the right hipbone. Place your finger about one and a half inches from the bony end. That is McBurney's point. Below that spot generally lies the base of the appendix. (A couple of people in 100 have theirs somewhere else in the abdominal cavity, one in 1,000 even on the left side!) Charles McBurney, the kind of brilliant, assured New York surgeon who was the obvious man to attend President McKinley as he lay dying from the assassin's bullet, made it clear in 1889 that the point was the location of maximal tenderness and pain, if indeed the patient suffered from an acute inflammation of the little finger-like pocket that protrudes from the end of the large intestine.

Appendicitis is neither rare nor new, as diseases go. Chimpanzees get it but monkeys do not. Hippocrates can be construed as noting such internal abscesses. The great Avicenna even remarked that, in cases of pain below the navel, it is a bad sign for the pain to remain in one place. Was this McBurney's point eight centuries ahead of the tireless surgeon from Columbia's College of Physicians and Surgeons? The little pocket somehow becomes obstructed (one man swallowed a .22 caliber bullet, and it was found neatly nested in his appendix), most often by abnormal dried pellets of fecal matter. Then its secretions cannot flow out, so that it swells, its blood supply is constricted, its cells degenerate and it "invites infection." An abscess sets in, or even the general infection of the entire lining of the abdominal cavity called peritonitis. If the patient is not treated, he has about a 50–50 chance of survival. Attacks as a rule recur.

In 1900 appendicitis was a cause of death about as frequent as automobile accidents are today. Its toll is down twentyfold, although it is still the cause of about half of all abdominal emergencies. More than half of the cases are seen in people between the ages of 15 and 30. This compact, engaging book, a narrative and a commentary alike, opens a surprisingly wide view on modern medicine and surgery through the special window of this one disease. Edward VII had his painful appendix successfully removed just before his coronation; in Britain and America that made the operation a symbol of smartness. It was only 16 years before that royal incision that the Boston pathologist Reginald Fitz had first recognized the inflamed appendix as the invariable starting point of perforations of the large intestine. He named the disease; he laid down the rules for fast action that still hold in acute attacks. Then surgical zeal (and the thirst for income?) went too far. George Bernard Shaw wrote *The Doctor's Dilemma* to satirize the craze for appendectomy without acute signs of disease, with the doctors blaming on "chronic appendicitis" just about all malaise. Most operations after such a diagnosis did not improve the patient at all, and by the 1930's this enthusiasm was at least deflected. "Chronic appendicitis is a doubtful clinical entity," modern texts say.

Antibiotics, intravenous fluids, intestinal rest (take nothing by mouth, so as to avoid peristaltic movements, which spread the infection) and ice packs can still cure the acutely ill. They remain the methods of choice whenever surgery under first-rate conditions is not available. The heroic pharmacist's mate who used flattened spoons as retractors and long-

nosed pliers as forceps to take out a buddy's appendix in the submarine *Grayback* submerged at sea in wartime 1942 succeeded. Nowadays that is against orders. In 1959 Vernon Houk was the only physician overwintering at the South Pole Station, and a young man developed the classical acute symptoms. The supply of local anesthetic had frozen and become unsafe. To use ether was too dangerous: there was a blazing fire in the surgery, and it was −90 degrees Fahrenheit outside. Houk chose the medical route. His patient did well, and when they parted at the end of the year, Houk gave the cheerful advice to "see your family doctor."

The appendix is not useless. It is a kind of abdominal tonsil: rich in lymphoid tissue, relatively large in childhood and somehow involved in the secretion of antibody. The author suggests very plausibly that the notable decline in the incidence of acute appendicitis in the U.S. is related to the dwindling performance of that routine tonsillectomy revered by middle-class therapists for decades. It is also suggestive that poor or rural populations with little medical care seem to have less appendicitis. That might be the result of diet, or it might be dubious statistics, or it might be the different history of challenge to and response by the immune system.

AN ILLUSTRATED FLORA OF THE NORTHERN UNITED STATES AND CANADA, by Nathaniel Lord Britton and the Honorable Addison Brown. Dover Publications, Inc. ($15 paperbound), and Peter Smith, 6 Lexington Avenue, Gloucester, Mass. ($22.50 hardbound). "The first complete Illustrated Flora published in this country" is here in a bargain three-volume facsimile of the 1913 edition, treating "every species, from the Ferns upward," growing wild from Newfoundland to the southern border of Virginia and westward to the western border of Kansas. There are almost 4,700 species, from the tiny duckweed to the tall white pine, 300-plus pages on grasses and sedges and more than 200 on thistles, each species with a handsome small-scale line drawing of the entire plant or of its key parts, each with a brief, technical botanical description. At a cost paperbound of less than .8 cent for every large, clear page, these volumes will fulfill any regional naturalist's need for the comprehensive: a classic reference of lasting usefulness, rich in common names for plants, although not at all a guide for the beginner. The same volumes are available clothbound as listed above, a good buy for hard users.

INDEX OF ADVERTISERS

AUGUST 1972

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

THE WANKEL ENGINE

THE NSU-WANKEL ROTATING COMBUSTION ENGINE. Walter G. Froede in *SAE Transactions,* Vol. 69, pages 179–193; 1961.

THE ROTATING COMBUSTION ENGINE. Roy T. Hurley and M. Bentele in *SAE Transactions,* Vol. 69, pages 641–649; 1961.

THE CURTISS-WRIGHT ROTATING COMBUSTION ENGINES TODAY. Charles Jones in *SAE Transactions,* Vol. 73, pages 127–147; 1965.

REDUCTION OF EMISSIONS FROM THE CURTISS-WRIGHT ROTATING COMBUSTION ENGINE WITH AN EXHAUST REACTOR. David E. Cole and Charles Jones in *SAE Transactions,* Vol. 79, Section 1, pages 205–215; 1970.

THE WANKEL ENGINE: DESIGN DEVELOPMENT APPLICATIONS. Jan P. Norbye. Chilton Book Company, 1971.

"IMPRINTING" IN A NATURAL LABORATORY

"IMPRINTING" IN ANIMALS. Eckhard H. Hess in *Scientific American,* Vol. 198, No. 3, pages 81–90; March, 1958.

IMPRINTING IN BIRDS. Eckhard H. Hess in *Science,* Vol. 146, No. 3648, pages 1128–1139; November 27, 1964.

INNATE FACTORS IN IMPRINTING. Eckhard H. Hess and Dorle B. Hess in *Psychonomic Science,* Vol. 14, No. 3, pages 129–130; February 10, 1969.

DEVELOPMENT OF SPECIES IDENTIFICATION IN BIRDS: AN INQUIRY INTO THE PRENATAL DETERMINANTS OF PERCEPTION. Gilbert Gottlieb. University of Chicago Press, 1971.

NATURAL HISTORY OF IMPRINTING. Eckhard H. Hess in *Integrative Events in Life Processes: Annals of the New York Academy of Sciences,* Vol. 193, in press.

THE NATURE OF AROMATIC MOLECULES

MOLECULAR ORBITAL THEORY FOR ORGANIC CHEMISTS. A. Streitwieser, Jr. Wiley-Interscience, 1961.

CARBOCYCLIC NON-BENZENOID AROMATIC COMPOUNDS. Douglas Lloyd. Elsevier Publishing Co., 1966.

ORGANIC REACTION MECHANISMS. Ronald Breslow. W. A. Benjamin, Inc., 1969.

AROMATICITY. P. J. Garratt. McGraw-Hill Book Company, 1971.

THE ANNULENES. Franz Sondheimer in *Accounts of Chemical Research,* Vol. 5, No. 3, pages 81–91; March, 1972.

THE BIRTH OF STARS

DARK NEBULA, GLOBULES AND PROTOSTARS. Beverly T. Lynds. University of Arizona Press, 1971.

THEORIES OF STAR FORMATION. D. McNally in *Reports on Progress in Physics,* Vol. 34, page 71; 1971.

LIGHT MOLECULES AND DARK CLOUDS. David Buhl in *Mercury,* in press.

MOLECULES IN THE GALACTIC ENVIRONMENT. M. A. Gordon and L. E. Snyder. Wiley-Interscience, in press.

DOCTOR-PATIENT COMMUNICATION

THE DOCTOR, HIS PATIENT AND THE ILLNESS. Michael Balint. International University Press, 1957.

THE MANAGEMENT OF THE DOCTOR-PATIENT RELATIONSHIP. Richard H. Blum. McGraw-Hill Book Company, Inc., 1960.

THE DOCTOR AND HIS PATIENT—A SOCIOLOGICAL INTERPRETATION. Samuel W. Bloom. Russell Sage Foundation, 1963.

GAPS IN DOCTOR-PATIENT COMMUNICATION, I: DOCTOR-PATIENT INTERACTION AND PATIENT SATISFACTION. Barbara M. Korsch, Ethel K. Gozzi and Vida Francis in *Pediatrics,* Vol. 42, No. 5, pages 855–871; November, 1968.

GAPS IN DOCTOR-PATIENT COMMUNICATION: PATIENTS' RESPONSE TO MEDICAL ADVICE. Vida Francis, Barbara M. Korsch and Marie J. Morris in *The New England Journal of Medicine,* Vol. 280, No. 10, pages 535–540; March 6, 1969.

GAPS IN DOCTOR-PATIENT COMMUNICATION: DOCTOR-PATIENT INTERACTION ANALYSIS. Barbara Freemon, Vida F. Negrete, Milton Davis and Barbara M. Korsch in *Pediatric Research,* Vol. 5, No. 7, pages 298–311; July, 1971.

ORIGINS OF THE BINARY CODE

TÉLÉGRAPHE MULTIPLE IMPRIMEUR DE M. BAUDOT in *Annales Télégraphiques,* Series 3, Vol. 6, pages 354–389; 1879.

LA TÉLÉGRAPHIE MULTIPLE in *Annales Télégraphiques,* Series 3, Vol. 22, pages 28–71, 152–177; 1895.

THE LAWS OF THOUGHT. George Boole. Dover Publications, Inc., 1953.

THE NEUROPHYSIOLOGY OF BINOCULAR VISION

THE NEURAL MECHANISM OF BINOCULAR DEPTH DISCRIMINATION. H. B. Barlow, C. Blakemore and J. D. Pettigrew in *Journal of Physiology,* Vol. 193, pages 327–342; 1967.

BINOCULAR INTERACTION ON SINGLE UNITS IN CAT STRIATE CORTEX: SIMULTANEOUS STIMULATION BY SINGLE MOVING SLIT WITH RECEPTIVE FIELDS IN CORRESPONDENCE. J. D. Pettigrew, T. Nikara and P. O. Bishop in *Experimental Brain Research,* Vol. 6, pages 391–410; 1968.

EYE DOMINANCE IN THE VISUAL CORTEX. Colin Blakemore and John D. Pettigrew in *Nature,* Vol. 225, No. 5231, pages 426–429; January 31, 1970.

FOUNDATIONS OF CYCLOPEAN PERCEPTION. Bela Julesz. University of Chicago Press, 1971.

CYCLIC AMP

CYCLIC ADENOSINE MONOPHOSPHATE IN BACTERIA. Ira Pastan and Robert Perlman in *Science,* Vol. 169, No. 3943, pages 339–344; July 24, 1970.

CYCLIC AMP. G. Alan Robison, Reginald W. Butcher and Earl W. Sutherland. Academic Press, 1971.

CYCLIC AMP AND CELL FUNCTION. Edited by G. Alan Robison, Gabriel G. Nahas and Lubos Triner in *Annals of the New York Academy of Sciences,* Vol. 185; December 3, 1971.

MATHEMATICAL GAMES

REFLECTED NUMBER SYSTEMS. Ivan Flores in *IRE Transactions on Electronic Computers,* EC-5, No. 2, pages 79–82; June, 1956.

AFFINE M-ARY GRAY CODES. Martin Cohn in *Information and Control,* Vol. 6, No. 1, pages 70–78; March, 1963.

DIGITAL TRANSMISSION OF ANALOG SIGNALS. William R. Bennett in *Introduction to Signal Transmission.* McGraw-Hill Book Company, Inc., 1970.

USING THE DECIMAL GRAY CODE. N. Darwood in *Electronic Engineering,* pages 28–29; February, 1972

THE AMATEUR SCIENTIST

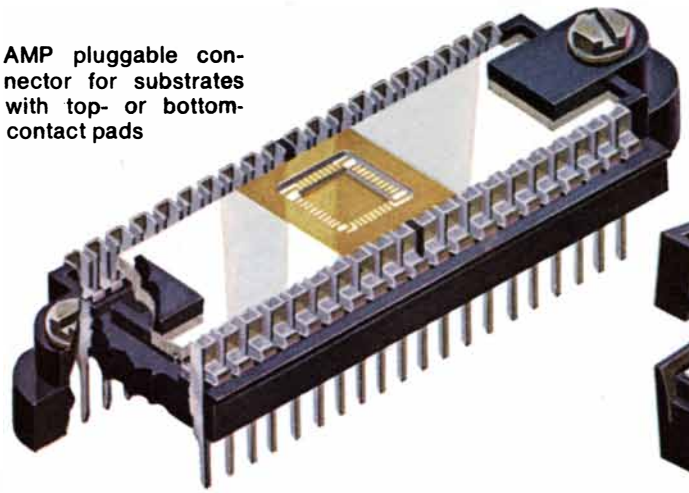
AMATEUR TELESCOPE MAKING: BOOKS ONE–THREE, edited by Albert G. Ingalls. Scientific American, Inc.

The Silver Martini.
For people who want a silver lining without the cloud.

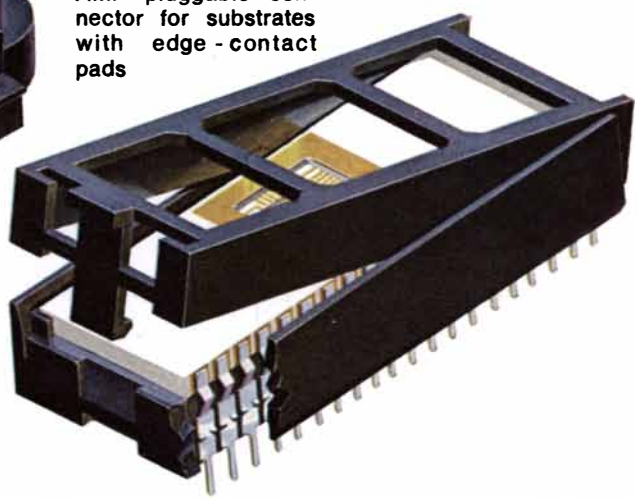
Smirnoff Silver
Ninety point four proof. Smirnoff leaves you breathless.

Now...practical pluggability for leadless LSI and LED packages.

AMP pluggable connector for substrates with top- or bottom-contact pads



AMP pluggable connector for substrates with edge-contact pads

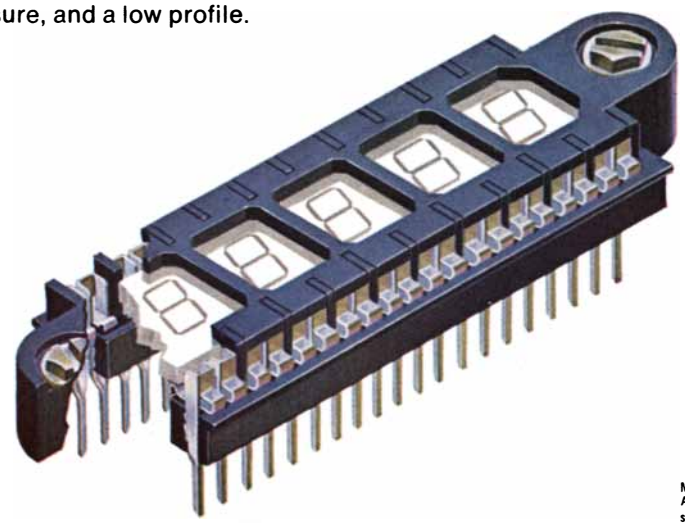


You've been hearing about the new leadless substrate LSI packages that will eliminate the handling problems of delicate leads and allow field replacement without mind-blowing soldering.

AMP now has a line of special connectors designed to make pluggability of these leadless substrates practical and reliable. The AMP connectors feature zero entry force insertion, a contact design with stored spring energy to maintain reliable contact pressure, and a low profile.

We have two types of these connectors for leadless LSI packages. One to mate with either top or bottom surface metalization, the other accepts the new packages with contact pads on the edge of the substrate.

Get the latest word on the latest in LSI/LED pluggability, by writing **Industrial Division, AMP Incorporated, Harrisburg, Pa. 17105.**



We also have a new version of this leadless connector that is designed to accept modular LED seven-segment displays. We can modify length and width to accommodate your complete LED display unit.

AMP INCORPORATED

Manufacturing and direct sales facilities worldwide: Barcelona, Brussels, Buenos Aires, Frankfurt, London, Mexico City, Paris, Puerto Rico, Sao Paulo, s'Hertogenbosch (Holland), Sydney, Stockholm, Tokyo, Toronto, Turin, Vienna.